



Measuring Trust for Crowdsourced Geographic Information

*A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Geographic Information
Science at the University of Canterbury*

Jeremy Severinsen, University of Canterbury
2015

Supervised by: Dr. Femke Reitsma
Dr. Mairead de Róiste

Acknowledgements

Four years ago I embarked upon a course of study that has stimulated, challenged and enthralled me, and would without doubt have been impossible without the support of a huge number of people.

Firstly, my thanks go to my supervisors, Dr. Femke Reitsma and Dr. Mairead de Roiste. From the early days of lectures and field trips, through to the last two year of thesis supervision, you have both challenged me to think outside the square, and reined me in when I strayed too far from it. Without your support this thesis would never have come about, and without your wise counsel it would not be finished in its current form. Your endless input, meetings, and late night Skype sessions are hugely appreciated.

Secondly, to my employer, Land Information New Zealand, who have sponsored this course of study and every element within. I have changed roles three times over the course of this study, and gratefully acknowledge the support of my direct managers, Nigel Smith and David Rhodes, as well as Geoff O'Malley of the NZ Geospatial Office. Particular thanks must go to the members of the Canterbury SDI team, for your enthusiasm for, and support of, the Building Our Footprints competition. Julian Carver, Stephen Bensberg and Becca MacGeorge – without you all this thesis could not have been finished. Furthermore, the guys at ECAN – Iain Campion and Ryan Elley – your technical prowess quite simply was the difference between a theoretical idea and an actual application. I am grateful to Ryan in particular for assistance translating my Python code into SQL. To my fellow students, Andrew Clouston and Trent Gulliver - we embarked on this journey together in 2010, and being able to share it with you has made all of the difference, on those long nights hunched over Wing IDE with telephone in hand. Further thanks to Paul Goodhue, Levi Mutambo, and others, for your help digitising footprints.

Finally, and most importantly, I would like to thank my family. To my boys, Finn and Oliver, thank you for putting up with all the time I spent at my desk, and more importantly, for taking me away from my work with tennis, soccer, swimming, video games, and Ben 10. You both constantly inspire me to be a better version of myself, and I hope that through this thesis I have inspired you towards higher education. I must thank Finn for his tireless efforts digitising building footprints, even with the knowledge that you could never win a prize. To Mel, I thank you from the bottom of my heart for always being there for me over these past years. Thank you for your guidance and suggestions, for being a sounding board for my sometimes indecipherable ideas, and for simultaneously giving me the active help and space that I needed to get it all done. On top of that you had your own work and study commitments, and of course our boys. It is your efforts that have truly inspired me to finish this thesis, and you were probably more integral to its completion than I.

Looking back at the magnitude of this support, and am both humbled and immensely grateful to you all.

Abstract

In recent years Crowdsourced, or Volunteered, Geographic Information (CGI, VGI), has emerged as a large, up-to-date and easily accessible data source. Primarily attributable to the rise of the Geoweb and widespread use of location enabled technologies, this environment of widespread innovation has repositioned the role of consumers of spatial information. Collaborative and participatory web environments have led to a democratisation of the global mapping process, and resulted in a paradigm shift to the consumer of geographic data also acting as a data producer.

With such a large and diverse group of participants actively mapping the globe, the resulting flood of information has become increasingly attractive to authoritative mapping agencies, in order to augment their own spatial data supply chains. The use of CGI would allow these agencies to undertake continuous improvement of their own data and products, adding a dimension of currency that has previously been unattainable due to high associated costs. CGI, however, through its diversity of authorship, presents a quality assurance risk to these agencies should it be included in their authoritative products. Until now, this risk has been insurmountable, with CGI remaining a “Pandora’s Box” which many agencies are reluctant to open.

This research presents an algorithmic model that overcomes these issues, by quantifying trust in CGI in order to assess its implied quality. Labeled “VGTrust”, this model assesses information about a data author, its spatial trust, as well as its temporal trust, in order to produce an overall metric that is easy to understand and interpret. The VGTrust model will allow mapping agencies to harness CGI to augment existing datasets, or create new ones, thereby facilitating a targeted quality assurance process and minimizing risk to authoritativeness.

This research proposes VGTrust in theory, on the basis of existing examinations of trust issues with CGI. Furthermore, a facilitated case study, “Building Our Footprints” is presented, where VGTrust is deployed to facilitate the capture of a building footprint dataset, the results of which revealing the veracity of the model as a measure to assess trust for these data. Finally, a data structure is proposed in the form of a “*geo-molecule*”, which allows the full spectrum of trust indicators to be stored a data structure at feature level, allowing the transitivity of this information to travel with each feature following creation.

By overcoming the trust issues inherent in CGI, this research will allow the integration of crowdsourced and authoritative data, thereby leveraging the power of the crowd for productive and innovative re-use.

Table of Contents

Acknowledgements	1
Abstract.....	2
Table of Contents	3
List of Figures	6
Acronyms & Abbreviations	8
Research Objective	9
1: Introduction	10
1.1 Background	10
1.2 Thesis Structure	12
1.3 Research Question	12
2: Literature Review	13
2.1 Trust in a Data Author.....	13
2.2 Spatial Trust - Accuracy & Precision	18
2.3 Temporal Trust – Issues of Provenance	23
2.4 Metadata & Data Structures.....	26
2.5 Summary	29
3: Methodology	31
3.1 Formulating a Generic Trust Model – The General Elements	31
3.1.1 The Spatial Components	33
3.1.2 The Temporal Components	33
3.1.3 The Total Model	34
3.2 VGTrust.py – Development of Code Package	35
3.3 Individual Model Components – Calibrating Case Specific Parameters.....	36
3.3.1 Calibrating Author Trust.....	37
3.3.2 Calibrating Spatial Trust	39
3.3.3 Calibrating Temporal Trust	41
3.3.4 Full VGTrust Model – “Building Our Footprints” Application	43
4: “Building Our Footprints” – A Facilitated Case Study	45
4.1 Background	45
4.2 Competition Implementation	46

4.2.1 Engagement Model for Participation.....	47
4.2.2 “Building Our Footprints” System Architecture.....	48
4.3 Quality Assurance Reference Data	50
4.4 Summary	51
5: Results	52
5.1 The Building Our Footprints Mapping Competition	52
5.2 Reference Data	52
5.3 Processing of the Raw Crowdsourced Data	53
5.4 Independent Quality Assessment	54
5.4.1 Object Quality	54
5.4.2 Statistical Data Cleaning.....	56
5.5 Analysis of the VGTrust Model	57
5.5.1 Attribute Correctness.....	57
5.5.2 Object Quality vs VGTrust	58
5.6 Exploring the Modelled Variables.....	61
5.6.1 Expertise.....	61
5.6.2 Experience.....	63
5.6.3 Activity Space	64
5.6.4 Spatial Precision	66
5.6.5 Temporal Quality	67
5.6.6 Relationship of Modelled Components to Each Other	68
5.8 Optimising the VGTrust Model	71
5.8.1 Attributes of the Data Author	71
5.8.2 Aspects of Spatial Trust & Temporal Trust	72
5.8.3 The Whole VGTrust Model	73
6: Discussion	74
6.1 Integrating Crowdsourced with Authoritative Data	74
6.2 Trust in the Data Author – Model Performance	76
6.3 Performance of Specific Spatial Indicators	82
6.4 The Effect of Data Provenance	86
6.5 VGTrust – Final Variable Weightings	87
6.6 An Effective Data Structure to Model Trust	88
6.7 Limitations of VGTrust	91
6.7.1 Limitations of the Base Trust Model.....	91
6.7.2 Omissions from the Base Model.....	93

6.8 Summary	94
7: Conclusion	96
7.1 Future Research Directions.....	97
7.2 Concluding Statements	98
Bibliography	100
Appendix 1: VGTrust.py.....	106
Appendix 2: VGTrust.sql	114



List of Figures

Figure 2.1 – The relationship between number of contributions to a crowdsourced feature and its quality, according to Linus’ Law.....	25
Figure 3.1 – VGTrust.py model. Extract from Python code module encompassing generic activity space calculation.....	35
Figure 3.2 - Process flow for the angle calculation component of the VGTrust model.....	40
Figure 3.3 – The relationship between number of contributions to a crowdsourced feature and its quality, according to Linus’ Law.....	42
Figure 4.1 – The information required to implement the VGTrust model was collected from participants at the time of registration.....	49
Figure 4.2 – Map image depicting building footprint extent, additional trust layer, shown as a green ticked box, and pop-up information window describing trust for that feature.....	50
Figure 5.1 - Feature capture at a school site. Participants have captured all features of interest, and have not been limited to buildings. Such features could include sports fields, parks, and recreational infrastructure such as tennis courts and swimming pools.....	53
Figure 5.2 - An illustrated example of Object Quality as defined by Zhan et al. (2005). The green illustrates the area of overlap between features in two different datasets, with the purple and yellow areas indicating the areas in each dataset respectively that did not intersect. This example depicts data from “Building Our Footprints” as compared to the reference data collected for this study.....	55
Figure 5.3 - Geometric error for a building feature. These errors appeared frequently in the dataset and were attributed to the unfamiliarity of users with the capture tool. These errors could easily be cleaned using automated business rules, and therefore present little risk to the viability of the VGTrust model. They were noted but not considered further in this analysis.....	56
Figure 5.4 - The sliver displayed above is the result of an overlap between two adjacent buildings in the same dataset. The extremely small area meant that the feature generated a high residual value, resulting in its exclusion from the subsequent analysis.....	57
Figure 5.5 - Results of a 2 - Sample t Test of Object Quality and VGTrust.....	59
Figure 5.6 - Comparative statistics showing the differences in distribution between VGTrust and OQ.....	60
Figure 5.7 - Relationship between expertise and Object Quality, where academic subject, or discipline, is linked to the spatial ability of the data author.....	62

Figure 5.8 - Relationship between experience and Object Quality, where year at school, or experience, is linked to the spatial ability of the data author.....	63
Figure 5.9 - Regression results showing the relationship between OQ and activity space, dissected into distance from school location and distance from home suburb.....	65
Figure 5.10 - Quadratic relationship between OQ and the number of vertices present, illustrating an increasing trend of OQ improvement while vertex count is between 4 and 20. A distinct deterioration of quality after this number is evident in the data.....	67
Figure 5.11 - Relationship between number of edits and Object Quality, where the maximum OQ value is unaffected by edit numbers, however the number of poorly scoring features markedly reduces with an increase in editing.....	68
Figure 5.12 - Principal Components Analysis for factor association in the modelled variables.....	69
Figure 5.13 - Response optimised regression for data author model components.....	71
Figure 5.14 - Response optimised regression for Spatial trust.....	73
Figure 6.1 - Heatmap showing density of footprint capture.....	81
Figure 6.2 – Capture scale issues illustrated.....	84
Figure 6.3 - Infographic depicting the components of VGTrust, with their weightings.....	88
Figure 6.4 - Overview diagram of VGTrust.....	95

Acronyms & Abbreviations

CGI	Crowdsourced Geographic Information
ECAN	Environment Canterbury Regional Council
GIS	Geographic Information System
LINZ	Land Information New Zealand
MGIS	Master's in Geographic Information Science
OSM	OpenStreetMap
SDI	Spatial Data Infrastructure
VGI	Volunteered Geographic Information
VGTrust	Trust model proposed by this research to infer quality for crowdsourced geographic information

Research Objective

Crowdsourcing represents a way to augment the spatial data supply chain of authoritative mapping agencies with a rich and abundant source of near-real-time information. The dramatic increase in the volume of geo-data available on the internet means that these data need simply be harvested and re-used to enhance traditional spatial products. The risk of course lies in its creation. Crowdsourced data is produced by a huge and diverse number of people, all with varying levels of expertise, experience, and motivations in the realm of spatial information – an environment that authoritative mapping agencies would quickly label “unknown” or “untrustworthy”.

This research seeks to bridge the gap between authoritative and crowdsourced data, to propose, test and establish an algorithmic model to verify the quality of crowdsourced geographic information. The model, labelled “VGTrust”, uses theories grounded in social network analysis, information about the author of that data, its provenance or history, as well as its uniquely spatial aspects, to infer a level of trustworthiness, or more accurately, to model its quality.

The objective of the model is to provide to authoritative mapping agencies a metric that will allow them to gauge the quality of crowdsourced information, in order to make appropriate decisions about incorporating these data into their supply chains. It will therefore allow a facilitated approach to crowdsourcing – improving the currency and relevance of authoritative data and products at a low cost, while at the same time reducing the risk associated with crowdsourcing. This risk is an inability of agencies to trust and rely upon crowdsourced data – an issue that it is proposed will be overcome by this research.

1: Introduction

1.1 Background

In recent years, the field of geographic information science (GIS) has become increasingly democratised, through a phenomenon that Goodchild has coined “neogeography” (Goodchild 2007). This phenomenon has drastically altered the role of the “user” of spatial information, as the proliferation of new, mobile, and spatially aware technologies such as smartphones and tablet devices has resulted in the ability for ordinary citizens to contribute local expertise to spatial datasets in volumes that are increasing exponentially (Seeger 2008). These consciously contributed data have been termed ‘Volunteered Geographic Information’ (VGI) (Goodchild 2007), and have been eliciting great interest from the scientific community. Given that this data is a rich source of near real time information, questions have emerged surrounding issues of integration with authoritative datasets and spatial data infrastructures (SDIs), hinged on how the quality of VGI can be accurately determined. This research uses the terms VGI and CGI (Crowdsourced Geographic Information) interchangeably, acknowledging the fact that while all data generated by crowdsourcing is CGI, “volunteered” geographic information could restrict the scope of this study unnecessarily to data that was specifically volunteered for a given purpose.

How then does a particular body of information gain a status of reputable? Is a piece of information recognised as inherently “better” than others based on the reputation of its creator, or is it when belief in a certain concept by many reaches such critical mass that a community accepts it as fact? There have been a large number of recent studies that have sought to address this “trust” issue with CGI, which have most often focussed on single measures of trust, and usually over the whole of a crowdsourced dataset, and often with comparison to existing authoritative datasets. Such trust measures can be broadly categorised into data author trust, spatial trust, and temporal trust in CGI, and it is these categories upon which this research is structured.

A large body of work including Golbeck et al. (2008) has proposed the idea of trust as a proxy for information quality, linked to the ongoing analyses of connections within social networks – if the contributor of a particular piece of data is deemed trustworthy, by a range of factors including reputation, then any data created by that trusted volunteer can immediately be deemed of a higher quality. An even larger body of work has focussed on the unique spatial quality indicators inherent in geographic information, and how these apply to the case of information contributed by a heterogeneous group of volunteers, often combined only through online user communities, and often influenced by a variety of motivations for contribution. Others still have posited the notion that the quality of a piece of data cannot be determined without reference to its temporal nature, or lineage and currency (Haklay et al. 2010, Aragón et al. 2009, de Longueville et al. 2010).

In 2007, Goodchild commented on the potential for geographic information collected by “the human side of the sensor revolution”. He stated that, “the six billion humans constantly moving about the planet collectively possess an incredibly rich store of knowledge about the surface of the Earth and its properties”, and that through the use of location aware and collaborative technologies (often referred to as “Web 2.0”), this knowledge can be translated into a rich and ongoing source of information, to be put to productive re-use under the same principles that guide interoperable data sharing inherent in any SDI (Goodchild 2007).

Since 2007, a significant body of research has been dedicated to exploring the means by which this prolific new data source could be integrated into traditional and emerging SDIs, both at a local and national level (Craglia 2007, Miranda et al. 2011, Budhathoki et al. 2008, Elwood 2008), which have traditionally been focussed on the sharing and freeing of data held by national mapping agencies. The tsunami of collaboratively created data rapidly approaching this island of authoritative agencies suddenly demonstrated the need for a paradigm shift to accommodate such an incursion, and the adaptation of elaborate standards to make these more accessible to the new “produser” class of neogeographers (Budhathoki et al. 2008, Goodchild 2008, Coleman et al. 2009)). These “produsers”, a new designation to define a group of people who actively produce geographic information at the same time as consuming it, subscribe to the norms of neogeography, and often also without any coordination or quality guidelines to which they adhere.

CGI datasets are created by such a heterogeneous authorship, that it becomes difficult to assess any of the quality indicators that are attached to more traditional forms of geographic information. Furthermore, CGI comes in many forms and does not adhere to a particular data structure, or groups of data structures, that coherently and actively capture all of the information required for a user to make an informed quality judgement on any particular feature or dataset. It is therefore important to establish not only trust for CGI, but a way to store trust information against these data.

Goodchild et al. (2007) proposed the idea of the “*geo-atom*”, the base form in which any piece of geographic data could exist. A “*geo-atom*” consists of a coordinate, or point, and an associated attribute value. In essence, this most basic data structure encompasses a large proportion of the CGI available on the Geospatial Web, and as such, this study seeks to propose a means of assessing trust in a geographic feature on the basis of author, spatial, and temporal components, and proposing an extended “*geo-molecule*” that will allow this trust information to travel in the basic data structure of each individual feature. Any potential re-user of this information could then use this quantifiable trust metric to determine the fitness for purpose of CGI for their given use, based on a quality threshold. For example, a national mapping agency, reputed for the provision of authoritative data (Johnson & Sieber 2013), could require a high quality threshold, while an ordinary citizen desiring a traffic update may be satisfied with a lower trust weighting to inform a decision that may have lesser consequences. Such an approach could facilitate the further

uptake of VGI by individuals, as well as larger agencies, and potentially infuse a widespread and current data source into a variety of decision making processes.

1.2 Thesis Structure

The following chapters of this thesis will explore the factors required to infer trust in CGI, and therefore quality, including a facilitated case study where the “VGTrust” model was deployed. The results of this study were analysed before a final “VGTrust” model is proposed. Chapter 2 will explore the existing literature that is germane to trust and CGI, and Chapter 3 will document the methodology used to construct a trust model from these parameters. Chapter 4 introduces a working case study utilising the “VGTrust” model, with the results of this implementation presented in Chapter 5. These results are discussed in detail through Chapter 6, before Chapter 7 offers thoughts for future research and a concluding assessment on the veracity of “VGTrust”.

1.3 Research Question

Fundamentally, this research will demonstrate a means by which CGI can be assessed for trustworthiness, and therefore inferred quality. “VGTrust” is a means by which authoritative mapping agencies can manage risk while at the same time augmenting their spatial data supply chains. Through the work of volunteers, the currency of these datasets can be drastically improved for little associated cost, by leveraging the power of the crowd. The purpose of this research is therefore:

“Can an algorithmic model be used to establish trust for CGI, thereby facilitating its assimilation into authoritative spatial datasets?”

2: Literature Review

There have been many studies presented regarding quality assessments for both CGI and VGI. Many of these have proposed theoretical solutions to the ambiguity surround trust in crowdsourced data, and some have illustrated practical solutions to aspects of these problems. This chapter will explore these studies, identifying aspects of CGI quality that can be carried forward and built into a trust model. The proposed trust model will incorporate the three broadly defined categories of quality assessment identified in the literature, and seeks to combine these into one holistic assessment of trust. These broad components are author trust, spatial trust, and temporal trust, which will be addressed separately in the following sections. This chapter will then explore the literature on data structures, in particular how this trust information can be stored and transferred with each feature.

2.1 Trust in a Data Author

A data author is simply defined as the person who has created any given piece of geographic data. When assessing CGI, given its often diverse authorship, certain characteristics of that person become important proxy measures for quality assessment. There are a multitude of personal aspects of a data author that are relevant to data quality, including their qualification, experience, and spatial ability. Trust in a Data Author can therefore be divided into sub-categories. In the case of this research, these are reputation and geographic proximity.

Expertise and experience are two factors intrinsically tied to the concept of reputation, which in turn is probably the most widely discussed across a range of academic disciplines. Reputation as a means to assess trust in a person has deeply established roots in social network analysis and wore widely on the semantic web. The terms “credibility”, “reputation” and “reliability” are often used interchangeably when assessing the source of a particular piece of crowdsourced data, and a large body of academic research has focussed on the idea that the credibility of a source of geographic information can be used as a proxy for the inherent quality of that data (Flanagin & Metzger 2008). This assumption hinges on several factors, including not only the definition of ‘credibility’ in the context of CGI, but also the nature and composition of the various networks and communities of producers (Coleman et al. 2009, Keßler & de Groot 2013) involved in the combined processes of data production and consumption.

The environment of Web 2.0 has produced an unbridled explosion of crowdsourced information, through social networking sites and other cloud based media outlets. These sites, although not specifically dedicated to the capture of geographic information, often contain a vast collection of georeferenced media and other data. Social networks are particularly germane to this research, not only because they contain a rich and ever expanding source of crowdsourced information, but equally because much research has already been dedicated to the assessment of trust and

credibility in these collaborative environments. Flanagin and Metzger (2008) define credibility as “the believability of a source of a message, which is made up of two primary dimensions: trustworthiness and expertise”. This definition has underpinned the majority of research in this domain, with Lankes (2008) concluding that the success of any collaborative (digital) environment is dependent on an appropriate measure of credibility. Flanagin and Metzger further argue that “assessing credibility inaccurately can have serious scientific, social, personal, educational, and even political consequences” (p159).

Mature crowdsourcing applications, such as OpenStreetMap (OSM) and Wikimapia, have implemented a quality assurance system based on the stratification of its user group, or “crowd”. The practice in essence applies the same principles used by authoritative mapping agencies by requiring a validation and approval process of new data to take place before these can be committed to the map. Despite the participants in the mapping crowd being volunteers, or ‘crowdsourcers’, the hierarchy of participation is stratified into differing levels, based on the performance and reputation of any given individual. All participants begin their mapping ‘careers’ at the bottom-most level, and can be elevated through the hierarchy according to the number and quality of their contributions, which contribute to their reputation, or trust, within the user community. This approach is consistent with the principles of trust and reputation on the semantic web outlined by Golbeck et al. (2008), and Golbeck and Hendler (2004).

A participant in this system therefore builds up reputation “credits”, in order to gain elevation within the system, and reflects the collaborative nature of CGI collection. It is interesting in that it seeks to essentially replicate the authoritative paradigm of quality assurance, only using assessors who most likely possess little formal qualification or expertise in data management, GIS, or mapping.

There have been several studies relating to the motivations of contributors within this system, such as those by Coleman et al. (2009) and Goodchild (2007). In reality, only a minimal number of contributors would possess either the ability, motivation, or drive to persevere long enough to attain this “master” status. Coleman et al. (2009), Elwood (2008), Heipke et al. (2010) and Haklay (2013) all discuss the characteristics of participation, the different types of people likely to participate in these projects and why, and identify that those people at the higher levels of trust within a project are necessarily the minority of very passionate mappers. These people could be passionate about the data itself, the particular project, or about the correctness and completeness of their local area. Whatever the reason, having a clear stratified minority group such as this promotes consistency within the dataset, and removes some of the potential bias that could be presented by a mapper with an ulterior or personal motive for providing a piece of data in a certain way. It also means that a collaborative mapping application is at a lower risk of malicious attack by disgruntled participants.

It therefore becomes essential to explore these concepts of trustworthiness and expertise in greater depth, in order to understand how they affect the changing nature of data provision following the rise of neogeographers, the participation in

geographic activities or generation of spatial products by everyday groups of people, often with no formal expertise in that area (Heipke 2010). An important distinction for this and Heipke's research is that, although relevant, establishing trust for passively collected crowdsourced data is not the stated objective, as fewer bias' are likely to exist in data not intended for reuse. It is the geographic information, volunteered to the collective group for an intended purpose, but produced by authors with little or no formal expertise in geography, surveying, or spatial sciences, where the assessment of trust and therefore value becomes important. Given the complex networked nature of the web, it is often difficult to assess the provenance of any data that has been generated – in fact with an indeterminable number of reuse cases, details about the origin of, and changes over time to, a dataset may simply be unattainable.

More facilitated examples of VGI, such as enormously successful Wikimapia and OpenStreetMap projects, attempt to overcome this issue by assessment of author credibility through a system of peer ratings and reviews, as well as recording the number and nature of edits made to any given feature (Mooney et al. 2010). While these factors ultimately impact on an author's quality rating, the user of CGI is not necessarily interested in obtaining data to a quality level usually associated with authoritative producers, such as government or professional mapping agencies. To this end, a user of CGI should assess the applicability of a source of data based on their own definition of fitness for purpose. What factors then are relevant for the assessment of 'fitness for purpose' as opposed to a more objective view of trust and quality? Or by establishing more objective measures of trust, can fitness for purpose then be established? It therefore becomes essential to return to the concepts of trustworthiness and expertise, and how the nature of these concepts in turn affect credibility.

In order to gain the most meaningful and transferrable assessment of CGI quality, effective measures of quality must be established at the most fundamental level for volunteered data, and only then extended and enhanced through a variety of additional measures. A large number of studies have identified that the majority of VGI is generated and collected in a collaborative environment, often involving information communities (Bishr & Mantelas 2008) who in and of themselves regulate the quality of volunteered features through a provenance process that Van Exel (2008) coined as 'Crowd Quality'. Essentially, these communities of users combine into formal or semi-formal entities not dissimilar to widely recognised social networks or professional networks, where individual connections can be graphed to illustrate the 'Small World Theory'. This theory has become more colloquially known as 'six degrees of separation' (Golbeck et al. 2008). Put simply, any two individuals can be connected through a graph of other mutually connected individuals, and that this connection can be established through a pathway of edges that includes no more than six other nodes (Golbeck 2008). As VGI user communities are most likely to be smaller and at a local scale, the Small World Theory becomes even smaller, with even fewer degrees of separation. The fundamental similarity to social networks remains however, and as such so does the relevance of interdisciplinary analyses of trust and reputation. Pickles (2011) further explores the fundamental psychology of collaborative web environments,

where communities of participants come together through the sharing of a common purpose. Described as “shared truth”, or “social dreaming”, the idea that online communities use the experience of others to augment and enhance their own, is a positive factor influencing participation of users, and increasingly relevant for future collaborations. Significantly, efforts to understand these factors on the semantic web are proving directly relevant to the analysis of CGI (Golbeck et al. 2008, Golbeck & Hendler 2004).

Golbeck et al. (2008) and Golbeck & Hendler (2004) present working case studies of a system known as “TrustBot”, which can be attached to email or messaging systems, and through a series of algorithms infer a trust rating for any given sender of electronic communicate. This assessment is based on specific paths, and the lengths of these paths, through a user’s ever expanding graph of personal connections. These ratings necessarily require a feedback loop of some kind (Gira et al. 2009) to appropriately assess relationships within the graph. Most critically these studies draw a clear distinction between the two fundamental elements of trust – credibility and expertise, and builds trust classes for each of these parameters. An overall *<trust_level>* subclass was built by assessing both credibility *<trustsPerson>*, and expertise *<trustsOnSubject>*. In essence, a person may be trusted as honest and reputable, however may not be deemed trustworthy on a particular subject due to lack of qualification or experience, also known as expertise.

Bishr and Mantelas (2008) identify five features of trust in the context of CGI. These are Transitivity, or the progression of trust through chains of people; Comporability, where different actors in a trust network similarly rate the same actor or item; Personalisation, or the more subjective component of a reputation rating – a feature of trust analysis that is in most cases the most difficult to quantify; Asymmetry, where trust between parties may not be equal in both directions; and spatial homophily, a factor unique to geographic information, where similarity of geometry, precision, and attribution are directly proportional to the inferred trust of that information object. This has been referred to by Haklay et al. (2010) as the Confirmation, or ‘Many Eyes’ Principle, and applied to analyses of the provenance of OpenStreetMap data. In particular, the analysis focussed on the number of rollbacks and edits of features which were used as a means to determine the quality and currency of that feature, and therefore its fitness for purpose and reuse. Interestingly, Haklay et al. (2010) also identify that a feature’s quality could be endorsed by a lack of change, with the “Many Eyes” principle being applied – the greater the number of views a feature has without any associated deletions or edits, then the accuracy and trustworthiness of that feature is directly proportional to that number of views.

Of course, as Elwood (2008) explains, “identity shapes knowledge and contribution”, and how then can an examination of trust and reliability properly ascribe a value to “the situational context in which the data is generated”? An examination of an author of information is required to tease out any motivations for contributions that may be fuelled by political or economic agendas, which may lead to data manipulations designed to invoke a particular reaction or achieve a

desired outcome. The motivations of this new class of data “producers” (Coleman et al. 2009) cannot be ignored. As with any non-authoritative source of information, objectivity cannot always be guaranteed, especially when the immense heterogeneity of VGI, and the inescapable fact that as “the human side of the sensor revolution” (Goodchild 2007), it becomes impossible to fully eliminate inherent bias’ from any given author. Without further investigation into the motivations of these contributors (Coleman et al. 2009), situated in a CGI context, trust on the semantic web can only be accurately quantified through one of its two inherent components – expertise.

Expertise is generally accepted to be the result of one or both of two factors – some form of formal qualification, or experience. In the case of neogeographers contributing to CGI datasets, qualification is in fact not formal, but linked to experience, particularly in a given area. Coleman et al. (2009) propose a spectrum of expertise for contributors, ranging from “neophyte” (someone with no formal background on a subject, but possessing the interest, time and willingness to offer their opinion (Coleman et al. 2009)) through to “expert authority”. Heipke et al. (2010) also seek to categorise contributors and infer expertise based on motivation, through designations such as ‘map lover’ and ‘casual mapper’, through to ‘experts’ and ‘open mappers’.

As previously discussed, CGI datasets are often driven by local user communities to inform citizens of, and provide solutions to, local issues. Goodchild (2007) explores these concepts further through the idea of an “activity space”, which is defined as, “the area within which the majority of an individual’s day-to-day activities are carried out.” Goodchild places a high level of trust in information gained from volunteers with a familiarity with their “activity space”, but acknowledges that this primary expertise is limited by temporality, and is valid only for the length of time that a person spends in any given activity space. De Longueville (2009) sought to evidence the truth of the ‘Activity Space’ theory, by facilitating the capture of a VGI dataset that included the provision of the author’s home location. This information was then analysed on the basis that the distance from that author’s home location to the encoded feature was inversely proportional to the spatial “degree of truth” of that feature. Interestingly, although this study did not actively collect information about the trustworthiness of the contributor, De Longueville et al. used information about the contributor to determine an assessment of a feature’s spatial quality, through an inferred user quality based on activity space.

Van Exel et al. (2008) identify user quality as a fundamental contributor to any assessment of trust in a volunteered feature or dataset, and a primary factor in their theoretical measure of trust labelled “Crowd Quality”. There is a key distinction for any quality assessment of VGI when compared to trust and reputational models based on social networks, as these omit factors that are unique to geographic information – the spatial and temporal components. For this reason, Bishr & Janowicz (2010) posit that reputation and expertise can only be assessed correctly in a VGI context through a measurement of information trust – whether a piece of data is fit for a particular purpose at a particular time, given that many of the traditional attributes and metadata will not be present for the user to make a fully

informed judgement. This “context deficit” (Flanagin and Metzger 2008), leads back to the predominant issue faced by many neogeographers – that author information has been identified in academia as a primary indicator of informational quality – yet this information cannot be guaranteed for a crowdsourced dataset.

There is a clear indication that reputational characteristics – namely expertise and experience - as well as the unique concept of Activity Space, are vital for any holistic assessment of trust in CGI. Additionally, recent studies such as Du et al. (2012), Haklay (2010) and Haklay et al. (2010) have augmented the focus of assessing credibility of source, to assessing credibility of information, in this case the uniquely spatial and temporal aspects of the data.

2.2 Spatial Trust - Accuracy & Precision

Trust assessment through data author is widely applicable across all types of crowdsourcing, however geographic data, by its very nature, includes a number of other components that need examining. A growing number of investigations are therefore beginning to question exactly what is special about spatial – in the case of CGI. While a large percentage of generic crowdsourced data, such as that created through social networking sites, is by this very fact attributable to a specific author, crowdsourced spatial information, or VGI, often contains very little source information. In addition to this, VG features contain an extra parameter that requires validation – its spatial accuracy and precision. While it is true that the potential quality of such a feature could be inferred by its author, in this case where that author was recognised as a geospatial or geographic expert, the simple fact is that this information is not always readily available, therefore other, more data specific measures must be investigated to determine the spatial quality of a volunteered feature.

As discussed in Section 2.1, Bishr and Mantelas (2008) proposed four parameters of trust within a particular user community, these being transitivity, comparability, personalisation and asymmetry, and also suggested a fifth parameter specific to VGI – spatial homophily. In essence, spatial homophily suggests that “similarity breeds trust” with the number of similar or identical features in a given location being directly proportional to the likelihood that those features will be more akin to their real world counterpart. Kuhn (2007) alluded to this issue in his discussion of uncertainty when data no longer stems from a single authoritative source. Indeed, there have been a significant number of studies conducted into the measurement and visualisation of spatial uncertainty, as well as several others offering comparison between volunteered features and traditional authoritative datasets. From these studies a comprehensive set of spatial quality parameters can be identified.

As existing facilitated VGI solutions have matured, so too have comparisons between the data collected through portals such as OSM and more traditionally authoritative sources. Rather than determine the intrinsic positional accuracy and

precision of crowdsourced features, OSM data has been compared to authoritative datasets such as the United Kingdom's Ordnance Survey, as in the United Kingdom OSM has achieved nearly complete coverage of the country (Du et al. 2012, Haklay et al. 2010, Craglia 2007, Osterman & Spinsanti 2011, Heipke 2010, Haklay 2010, Keßler & de Groot 2013, Fan et al. 2014).

Haklay (2010) identified eight factors essential for determining the quality of geographic information, volunteered or otherwise, previously proposed by Van Oort. These factors are summarised as Lineage, Positional accuracy, Attribute Accuracy, Logical Consistency, Completeness, Semantic Accuracy, Usage, Purpose and Constraints, as well as Temporal Quality. It is well known that by its very nature, there can be no guarantee that a volunteered dataset will be complete. With no top down coordinated approach to ensure full coverage, a volunteered dataset relies on the “self-organising capacity of crowdsourcing ecosystems” (van Exel et al. 2008), which essentially means that full coverage by a dataset will only be achieved once a certain threshold is reached with regard to volunteer numbers. Such a situation has recently been realised in the case of OSM, and Haklay et al. (2010) cite that OSM road data is comparable with traditional authoritative with regard to both positional accuracy and completeness, although this is not true for all feature types. A well-travelled road network is significantly easier for volunteers to map than other, more abstract or remote feature types. Further research is required into the assessment of completeness of a dataset in a volunteered context, and many authors continue to refer to the ‘digital divide’ (Heipke et al. (2010) as a phenomenon that will remain a barrier to any attempt to map the world through volunteers. Heipke summarises this phenomenon clearly – volunteering geographic information requires access to both affordable technology and the internet – pre-requisites that are simply not available in a global context, and will even differ significantly between socio-economic groups and regions within any one city. Indeed, feature richness within any such dataset is often directly proportional to the relative affluence of potential participants within its subject regions, and as a result, analysis of completeness as a quality factor for VGI should be subject to further, and more specific, studies. This investigation is focussed on assessing the quality of VGI at a feature level, for which positional accuracy as a measure of quality bears more relevance.

Completeness of a dataset is key in these studies, which aggregated CGI features within a given buffer of an authoritative feature. Much study has been dedicated to the assessment of road networks in the UK (Haklay 2010, Du et al. 2012). A significant number of contributions are GPS logs of people's transit through the road network in the UK. Depending on the spatial precision of their device, the triangulation of position by cell towers or the position of navigation satellites at a certain time of day, two contributors travelling on the same course could have differing data by up to several metres. These studies seek to aggregate these data within a specific buffer from their authoritative counterpart, and therefore draw conclusions about the quality of that data.

There have been several further studies seeking to assess the positional accuracy of VG through comparison with authoritative datasets. A recent study by Du et al.

(2012) attempted to identify methods to facilitate the integration of crowdsourced and authoritative datasets. Although this approach focussed heavily on methods of data cleaning, Du et al. made use of several techniques that can be applied to a standalone crowdsourced dataset in order to address issues of feature vagueness. By making use of what they described as a “fuzzy distance” through the implementation of a feature buffer, Du et al. determined which volunteered features lay within a tolerable distance of the authoritative example, and could therefore be considered suitable for integration with that feature. This approach is not dissimilar to that applied by Haklay (2010) and Haklay et al. (2010), where percentage intersections with buffered authoritative features were used to determine the relative positional accuracy of volunteered features sourced from OpenStreetMap.

De Longueville et al. (2009) identified several additional attributes for assessment of spatial precision, and extended the concept of buffer use as a means to visualise the vagueness rating computed for any given feature. Having described geographic vagueness as “any attribute that does not conform to Boolean logic”, de Longueville et al. structure their research around the concept of “degrees of truth”, a concept that this thesis seeks to assess through the combination of a variety of relevant quality indicators, and further links to the “fuzzy sets” or “fuzzy distance” theory previously discussed with reference to Haklay’s work. An interesting assessment of geographic vagueness is undertaken by De Longueville et al. (2009), where fuzzy levels of geographic precision are likened to an “egg yolk” and “egg white”. The ‘yolk’ represents the certain aspects of a features location, whereas the ‘white’ represents elements of uncertainty. This method is valuable when aggregating large numbers of crowdsourced features into one representation. When all features of a certain type are aggregated, the cluster in the centre becomes the ‘yolk’, where there is certainty of position. Additional outliers become the ‘white’, where there is a lesser number of contributions, and thus a lesser degree of certainty, or geographic vagueness.

Following from the notion that the collection of CGI is, in every case, heterogeneous, no two independently derived features will be precisely in the same position. While a VG dataset may be incomplete in some geographic areas, in others the richness of information may result in a significant duplication of features, providing a barrier to reliable reuse of that dataset. How then does a re-user of CGI determine if two separate entries in a geographic database are in fact duplicate representations of the same real world object? The topological implications are obvious, and sometimes seemingly unsolvable without consistent attribute information. Several studies have introduced the concept of crowd endorsement, including van Exel et al. (2008) and Bishr and Mantelas (2008), who describe crowd endorsement as the “harnessing of the collective intelligence of information communities”. Simply put, the greater number of people contributing similar features at a given location, the more likely it will be that that particular location, shape, and attributes of that feature are a correct representation of the physical world. Bishr and Mantelas (2008) describe this concept as “spatial homophily”, or that similarity breeds trust. Du et al., Haklay et al., and Haklay have all employed the use of buffers to facilitate the “fuzzy distance” theory of feature integration,

essentially determining the most likely single position of a feature based on the proportion of features within a certain threshold that purport to be representing the same real world object. This approach seeks to resolve logical consistency issues and optimise reuse, but further requires an increased formalisation of VGI collection, and consistency of post processing if VGI is to be considered appropriate for integration into authoritative datasets (Johnson and Sieber 2012).

Furthermore, de Longueville et al. identify several additional factors which they consider in their assessment of spatial precision. These factors do not consistently appear in later studies such as those of Haklay, most likely because of the facilitated nature of the de Longueville research. While Haklay (2010), Haklay et al. (2010) and Heipke (2010) attempt to establish measures of positional accuracy for genuinely crowdsourced features, de Longueville et al. have approached the issue from the standpoint of facilitated CGI, through which a range of additional parameters can be collected implicitly and explicitly from every volunteer. These additional parameters capture details about the scale at which a feature was created, the scale at which it was modified, and the contributor's own assessment of how vague they consider their capture to be. The latter of these has little relevance for non-facilitated CGI, as there would often be no opportunity of motivation for a volunteer to provide critical feedback on their own contributions.

Such an approach also presumes that author information is collected, and author reputation established, which are not guaranteed outcomes across the wider field of CGI. On the other hand, information about scale at time of capture is a factor that could be considered more widely in the CGI community, for even the most basic of volunteered geographic features. In other words, the scale at which a feature is created or modified is tied to the expected level of precision for that feature. The more familiar an author is with a created feature's real world counterpart, the more likely that contributor will be to create the feature at a smaller scale, with a higher level of detail and spatial precision. Therefore by collecting scale information at capture, a subsequent user of that feature could infer judgement on the author's knowledge and expertise, and thus the extent upon which the spatial accuracy and precision of the feature could be relied (De Longueville et al. 2009).

Interestingly, de Longueville et al. consider this information to be metadata as related to any given feature, and store it in a data structure that they associate with author testimonial. It could be argued that this information, due to its widespread availability, could be stored as a fundamental attribute of any volunteered feature, and incorporated into basic data structure for crowdsourced geographic information. This approach, of course, applies directly to features *digitised* through an online portal, and does not account for those features captured in the field using location aware devices, such as GPS units, mobile phones, tablet computers, and navigation systems. Although not specifically mentioned in their work, the rationale provided by de Longueville et al. for inferring quality from scale could logically be extended to infer that any feature captured first-hand by an instrument in the field. Such in-situ capture could be awarded the highest quality rating for scale, as it is

simply not possible to view any real world phenomenon with as much detail and precision as when viewing first hand.

The choice of instrument in and of itself presents additional uncertainties when assessing the accuracy and precision of volunteered features. While authoritative datasets such as national cadastres consistently contain metadata about how any given feature was captured, CGI for the most part excludes any reference to this. As the accuracy capabilities of consumer devices increase exponentially over an exponentially shrinking timeline, the “device issue” as articulated by Goodchild (2008), Craglia (2007) van Exel et al. (2008), Mooney et al. (2010) and others has given way to a more general “completeness issue”. Historically, device metadata was used to distinguish between information that could be considered of “professional grade”, such as cadastral work undertaken by surveyors, using higher grade equipment to deliver more precise results, with that information collected by the community of neogeographers, who could potentially collect geographic features using a range of devices of varying degrees of precision, from phone to handheld GPS units.

Compounding this issue is the fact that any volunteered information is just that – volunteered. Van Exel et al. (2009) articulate that there is no onus on a volunteer to identify instrument information when volunteering a feature to a crowdsourced database, and furthermore, the precision of any given feature, by means of logical consistency, can have its own spatial precision affected by the likely unknown precision of its neighbouring features. There is often uncertainty about whether two spatial features are in fact representations of the same physical place Goodchild (2008).

This problem is not limited to choice of instrument. Heipke (2010) and Goodchild (2008) signal issues with base imagery and context data as fundamental contributors to vagueness associated with CGI. Heipke identifies that the means by which crowdsourced geographic features are created can be aggregated to two main categories – those created from location aware sensors (predominantly GPS), and those digitised from orthorectified imagery.

It is important that any CGI application provides its volunteers with context data and imagery that can aid them in providing high quality information (Seeger 2008). With a significant proportion of all CGI being created in a web-enabled desktop environment, the case for quality metadata on the context information is given more weight. It is important in this case to draw again the distinction between crowdsourced geographic information and that which is actively “volunteered” in the truest sense of this paradigm. Geographic information that is collected by, to adapt Goodchild’s term, “citizens as *involuntary* sensors”, is by its very nature collected in the place of its real space-time measured object, and thus remains unaffected by the quality considerations attached to base mapping and other context data.

The risk associated with base imagery centres on its context deficit. In most cases, imagery providers do not provide “even the most obvious elements of data quality,

the date and time at which the base imagery was acquired, and its spatial resolution” (Goodchild 2008). Goodchild cites further examples of significant misrepresentation of imagery, where comparisons between providers in one case exhibited an approximate fifteen metre discrepancy between base imagery services (Goodchild 2008). Given that the vast majority of volunteered features are of an urban nature and related to objects and phenomena at a streetscape scale, the implications of this could potentially result in a feature captured from one misregistered base image showing as on the opposite side of the street if reused with another. This influence on spatial quality and precision is one of the most difficult to quantify into a single metric. As Goodchild points out, visual superimposition of various base images and volunteered features show a clear lineage (Van Exel et al. 2008, Goodchild 2008) of which features were produced from which context data. Du et al. (2012) comment on the effect of this on the logical consistency or topology characteristics of a dataset, although in facilitated situations where single VGI datasets are set in context of a single base image, the potential of this issue is negligible. Goodchild (2008) also proposed that any assessment of fitness for purpose of a CGI dataset should include information on which base map was used in the creation of the data. Such a concept is supported by Poole and Wolf (2012) in their discussion on the future use of metadata for CGI, and could include information on any base mapping service with which a crowdsourced geographic dataset is positionally compatible.

Of course, this approach directly contradicts many of the OGC standards and quality indicators explicitly established to facilitate widespread interoperability of *datasets*. Even so, the majority of investigations into this ever expanding domain agree that there can be no escaping the heterogeneity of CGI, and that each quality or fitness for purpose assessment of a crowdsourced dataset must be considered on a local level using local factors, as opposed to the global alternative (Seeger 2008, Heipke 2010, van Exel 2008, Goodchild 2008). As such, perhaps positional accuracy and precision can only be considered as relative to any range of datasets and any point in time – imagery can be re-registered, and official projections can change – and in light of this, if there remains a desire to coordinate interoperability between such datasets, then this could be a descriptive role of a future dynamic standard, although the costs of maintaining this relative to the potential benefits of CGI are yet to be fully investigated.

2.3 Temporal Trust – Issues of Provenance

It is clear that change over time is an important concept to consider when investigating the precision and accuracy of a crowdsourced feature. Rapid change to a feature can signal that it is dynamic in the real world, confirming CGI as a valuable source of up to date data. It is, however, important to consider that not all changes over time to a feature’s geometry or attribution are the result of poor quality of the original feature, or that a lack of change means that the original data is any less accurate. A real world object may not have changed either. With this in mind, it is clear that in order to properly assess a quality metric for VGI based on

author reputation and indicators of spatial precision, any assessment must also consider a third dimension of trust – temporal quality.

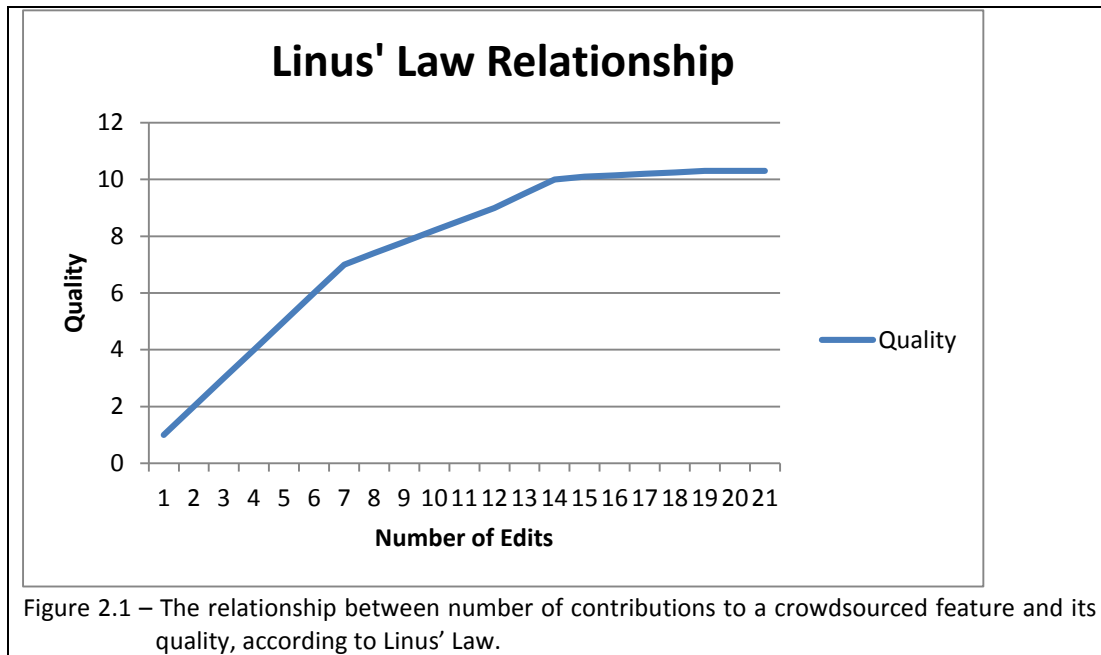
The majority of current GIS data models rely on the spatial definition of a feature as the primary source of identity for an object (Goodchild 2008). In the case of VGI, this means of orienting any database is limiting and scattered with risks, due to the fact that the geometry of any feature could change with such frequency that any assessment of a feature's provenance and lineage could become difficult to the point that it becomes prohibitive to re-use. As the geometry of a feature could be the subject of ongoing change, then identity must be maintained by one of the other attributes, although gives no indication of which attribute should take on this role.

The lineage of a dataset becomes particularly important when it is considered that the “up-to-date” nature of CGI forms the bulk of its appeal, not only at an organisational or fundamental data level, but also at the level of an individual feature. Kuhn (2007) stated that “with the vastly increased, often near real-time availability of spatially referenced information, analysis capabilities grow significantly”, and it is this feature of CGI that has proven attractive to both citizen communities and larger organisations, with an ever active sensor network providing information with a regularity and timeliness that would simply be impossible for any national mapping agency to achieve. The cost of this benefit is that traditional quality considerations are omitted in favour of currency. Du et al. (2012) highlighted currency as one of the primary indicators of data quality, along with consistency, accuracy, richness of information and fitness for purpose. Often, re-users of CGI are simply interested in the most up to date information, such as real time traffic congestion, however the logical extension of capturing information about the lineage of each feature within a dataset, is the ability for much greater analysis of processes and changes over time in any given community.

A number of studies make use of OSM data to explore this lineage, and to therefore make inferences about the quality of those features (Mooney et al. 2010, Haklay 2010, Du et al. 2012). Of course even CGI in its most basic form is time-stamped at its creation (Aragó et al. 2009), but of particular use in the case of OSM is that all edits are logged and can be tracked with associated timestamps. There have been two approaches to considering the panoply of changes that can affect a feature at any stage after its inception. The first, as considered by Haklay et al. (2010) is that each amendment can be considered a *refinement* of the original feature, and that the associated improvement in overall quality is a reflection of the applicability of Linus' Law. The second approach, as defined by Mooney et al. (2010) and refined by Trame and Keßler (2011), interprets each time-stamped amendment to a feature as a reflection of a real world change to the object represented, and as such, captures and allows modelling of change over time.

Linus' Law can be summarised as a direct but non-linear correlation between the number of contributors to a feature and the quality of that feature. Haklay et al. (2010) disproved the presence of an overall direct linear relationship between the number of contributors and feature quality, instead identifying that the first six

contributions made to a feature have the most influence on its positional accuracy and attribute correctness (Figure 2.1). Once the number of contributors to a feature exceeds fifteen, there becomes almost no discernible impact on quality. This approach was tested using by comparing volunteered road network data with a variety of traditionally authoritative sources, and comparing this with the number of contributors within a given area unit, with the express goal of proving Linus' law as a means to assess trust and quality on a standalone basis (Haklay et al. 2010).



Aragó et al. (2009) propose two formulae for assessing another theory on temporal quality – that no change over time may represent no change in the real world, or in fact an endorsement of the quality of the original feature. The first they label *Change Ratio*, which assesses the number of changes made to a feature in the time between its creation and final edit. The second they describe as *Contribution Ratio*, which seeks to measure the total number of both changes and endorsements to a feature within that same time period. Both ratios account for times a feature has been viewed but not changed, and therefore endorsed.

These representations of process can serve to identify areas of rapid change, such as a new city subdivision, or conversely show areas where little change is evident, such as a protected historic place. Roick et al. (2012) posit that by analysis of these temporal processes, a user of CGI can track the evolving activity of its contributors, and identify community areas of interest, or areas that require more attention from volunteers in order to achieve completeness and logical consistency within a dataset.

Ye et al. (2012) suggest using timestamps to classify features and explore semantic inferences using user check in's at points of interest. Ye et al. argue that it is not simply physical space that defines any given feature, as that space may change in its use and purpose at particular points in its lifecycle. For example, during the day

or business hours, a particular establishment may be classified as a cafe or restaurant, however as time progresses into evening, this could more appropriately be considered as a bar or nightclub. Ye et al. have described this approach as the “*temporal – semantic interaction*”, and should be the subject of further research in this area.

Fundamentally however, the temporal value added by CGI is its currency. CGI has been employed in several coordinated responses to natural disasters due to the speed at which up to date information can be collected and disseminated (Zook et al. 2010, Poser et al. 2010), therefore reinforcing the information relevance. It is essential to not only capture information about a feature’s creation and most recent edition, but to also place this into the context of general change in its immediate area – due to a lack of change to a real world object, a feature with an aging creation date may still accurately reflect that object, and therefore remain current and relevant. A feature may also be one of many, although slightly separated, instances of the same real world object, created by a number of users. Parker et al. (2012) state that CGI is likely to be most relevant to the user when a geographic feature is dynamic rather than static in nature. Therefore any quality assessment of the temporality of VGI must therefore account for currency in the context of its surrounding features and general surrounding activity. Keßler & de Groot (2013) reinforce the value of CGI as a way to augment and enhance authoritative datasets, where traditional mapping agencies lack the resourcing to keep their data and products both of a high quality and up to date.

2.4 Metadata & Data Structures

CGI is inherently heterogeneous, which often creates a barrier to data discovery. These data have historically been treated by business and academia as fascinating, although being of questionable reliability and, therefore, questionable use. It can be argued that even at its most basic level, a CGI feature can contain information that will determine to some extent how it can be trusted.

Goodchild et al. posited the idea of the “Geo-atom” – the most basic and fundamental building block of geospatial data (Goodchild et al. 2007), a structure which applies to a vast collection of crowdsourced data. At its most basic level, a piece of data could consist of a coordinate – a point feature – with an associated attribute of some value, usually of a thematic nature. An example of this could be a name of a restaurant associated with a specific location. At its most fundamental level, a geo-atom can therefore be described as a location, and some value associated with that location. Can such a simple data structure be considered to hold sufficient information to determine its trust, and therefore the reliability of that feature? While some measure of credibility may be established from these basic attributes, the question persists about which other fields may be required to change a *geo-atom* into a *geo-molecule* that can appropriately and reliably describe trust.

In contrast to such a simple data structure, authoritative datasets produced by mapping organisations, contain a significant amount of metadata. These organisations have by their very nature and history built a level of trust in their products. As previously discussed, these organisations were traditionally responsible for producing geospatial datasets when the cost to do so was prohibitive for neogeographers (Johnson & Sieber 2013). Authoritative datasets are often armed with a large and complex body of metadata, which aids discoverability by adhering to voluminous rigours of international standardisation. Here it is argued that all of the information needed to determine trust can be found in the metadata of a dataset, although this somewhat blinkered approach seems linked to the notion that a dataset does in fact have associated metadata, and therefore was produced by a professional body capable of enforcing quality standards. The average data volunteer is equipped with no such skillset, with Poore and Wolf (2013) noting a blog post that stated, “unless a caveman can do it, users won’t read or write meaningful metadata, and relevant metadata must be stored and travel with the data”. Such a mind-set raises a particularly relevant question for CGI which is, to what extent should information normally regarded as metadata be stored within the actual data structure of a dataset? In the case of CGI, given that each individual feature could realistically be the product of a different author, does this information require capture and storage at the base data structure level?

Sui, Goodchild, and Elwood (2013) describe CGI in the context of the exaflood – that is the exponential tsunami of user produced information in the past five years – and situate this in the frame of ease of user interaction. They argue that much effort has been dedicated to improving geoportals that interact with a user, making these more intuitive and simple. An example is the widespread popularity of Google Maps and its associated API, an interface which has become synonymous with personal navigation amongst the user community. On the other hand, data structures and metadata have become increasingly complex, and therefore unmanageable by neogeographers. The same can be said of geographic information in the world of Web 2.0 – data and metadata structures have become complicated to a level that discourages their use by the crowd. As a result, and since the populating of these fields is not a mandatory function of creating geographic data, many crowdsourced datasets simply lack this information. Sui et al. further note that there exists an unusual paradox that metadata must simultaneously be more simple as well as more comprehensive, and that there is a clear need for the metadata to follow the feature (Sui et al. 2013).

In the context of assessing the trust level of CGI, this phenomenon must be first examined at its most fundamental level, without the additional parameters and metadata generated by more facilitated VGI systems, such as OSM, Wikimapia or other established platforms). Whether a particular feature can be classed as crowdsourced or volunteered, when left unchecked to proliferate in the Geoweb, there are only a handful of parameters that can be assured for any given user generated feature. At its most basic level, a volunteered feature would consist of a geographic coordinate – a point – and some form of attribute. From this, further point instances can be combined to construct lines, polylines, and polygons, and these features could theoretically infinite collection of other attributes. These

notions are return to the concept of the “geo-atom”, where Goodchild et al. (2007) described the most basic building block of geographic data as:

$$\langle x, Z, z(x) \rangle$$

“where x defines a point in space–time, Z identifies a property, and $z(x)$ defines the particular value of the property at that point” (Goodchild et al. 2007 p.243).

Given that assessing trust in a volunteered feature has been regularly linked to both author and spatial elements (Section 2.1 & Section 2.2), and that such a geo-atom contains no information about the source or provenance of the data, is there sufficient information in such a basic data structure for a user to make a determination about the fitness for purpose of such a feature?

In a facilitated environment, several additional attributes can be collected to make inferences about trust. This most basic data structure cannot account for such parameters, however, and as such a user is left with very few avenues down which to explore notions of quality. A possible solution could be through crowd-consensus, where the volume of similar contributions, in both space, time and content, can be seen as crowd endorsement for a particular feature (see Section 2.2). This approach is not without its caveats, however, as crowd-endorsement can logically lead to a group mentality and crowd consciousness, where inaccuracies and bias’ are perpetuated by what could be considered a “pack mentality” (Gira et al. 2012). Van Exel et al. (2011) incorporate the idea of consensus into their theory of “Crowd-Quality” as a measurement for CGI, and further argue that some measure of currency is essential to determine whether a particular feature has been the subject of group-bias.

Consequently, a volunteered geo-atom may be able to provide a rudimentary assessment of its own quality, and this may be sufficient for a user’s particular purpose, however for the majority of users who have very little expertise in determining trust for data, or for large scale geographic data producers who may wish to supplement their authoritative datasets with CGI, the geo-atom simply cannot meet the required quality threshold. Furthermore, with this notion established, what data descriptors are therefore required to adequately determine trust to a particular standard, so that it can be assessed as “fit” for a variety of purposes.

The fundamental means of determining fitness for purpose has always been metadata. Simply described as “the data about the data”, metadata has evolved as a part of the top-down paradigm of geospatial information dissemination (Poore & Wolf 2013). Poore and Wolf identify that the creation of metadata is not only onerous for professional organisations or trained geospatial professionals, but even more difficult for those “producers” (Coleman et al. 2009) who are emerging in great numbers and publishing geographic data in the distributed world of Web 2.0. The result is that these vital details are often omitted from single features as well as whole datasets.

With the changing environment generated by the Geoweb, professionals and volunteers alike have begun to call for a more user-centric approach to the creation and capture of metadata, although perceptions differ as to how this can be achieved. Neogeographers generally advocate for a simpler metadata structure, while professionals and academics see the solution to discoverability problems as lying in the capture of additional information as metadata (Poore & Wolf 2013). This is the heart of the apparent metadata paradox, which in turn is driven by the four areas where metadata needs to change in order to keep pace with the results of a Web 2.0 society. These are usability, support for co-production of data by communities of users, findability, and the relationship between data and metadata. Indeed there have been a number of studies in recent years that attempt to capture aspects of VGI quality, including uncertainty, however there remains no consensus on a data structure that will provide sufficient information to generate a trust metric. Goodchild (2007) has drawn attention to the fact that the traditional lines between data and metadata have become increasingly blurred, and further stated in 2009 that,

“developments in the Geospatial Web have leapt far ahead of any concern for confidence limits of metadata, so information on the uncertainty associated with locations is almost certainly unavailable, and unlikely to be inferred...”

Indeed, studies attempting to clear these murky waters, such as a 2009 geographic vagueness study by Longueville et al., resulted in a complex table structure that stored information about the creation of a feature, including the user who created it. Unusually, this study associated data about geographic precision or vagueness against the creator of the feature, as opposed to against the feature itself. Although this approach proved effective in calculating a degree of truth for each feature (Longueville et al. 2009), it raises questions about the ongoing clarity of quality for these features, if this dataset was to be discovered and used by others for differing purposes. Fundamentally, any data structure employed to capture information about trust must solve the Poore and Wolf quandary – it must capture additional information that is not currently present in the “geo-atom” version of crowdsourced data, but at the same time make this information easy to understand and a more appropriate way to inform re-use and fitness for purpose.

2.5 Summary

There are an array of descriptors for trust in CGI, this chapter having separated them into three broad categories – Author Trust, Spatial Trust, and Temporal Trust. An examination of existing studies reveals that despite this separation, and the fact that the variables have been addressed separately in previous analyses, they retain an inherent connection to one another. This chapter has also identified that a comprehensive assessment of trust can be difficult to finalise for pure crowdsourcing applications, and the best way to achieve confidence and trust in CGI is through facilitated solutions. Chapter 3 – Methodology will therefore focus

on the development of a holistic model, accounting for the factors and principles described in this chapter.

There is a need therefore to find a holistic solution that will enable re-users of CGI to anticipate the reliability of these data. A means to explore trust through a combination of descriptors has yet to be proposed by the industry. With this in mind, the model proposed by this thesis will combine author, spatial and temporal trust for a more robust assessment of CGI. Furthermore, this thesis proposes a data structure that, through an extension of the geo-atom, will allow all of the metadata needed to determine trust to carry with each individual feature.

The model, labelled “VGTrust”, will be specifically targeted at authoritative mapping agencies, to enable trust in CGI and allow its integration into existing authoritative datasets, and will produce an easy to understand metric on the basis of the principles described in this chapter.

3: Methodology

This research will propose a model for assessing trust in crowdsourced geographic information. This model is based on the individual components identified in relevant literature, and is labelled “VGTrust”. The components can be broadly categorised into three groups –the author of that feature, various unique spatial elements, and temporal components of geographic data quality. This chapter discusses the approach used in this thesis to formulate a generic and transferable model from these component factors, at an appropriate weighting, and describes how VGTrust was calibrated for one specific implementation.

In theory, this model is generic - each implementation in a crowdsourcing environment will require a case specific calibration of the series of trust factors, with major differences existing between crowdsourcing undertaken by applications for digitising, versus field collection using mobile devices. The VGTrust model as a single unit is a novel concept and no known dataset existed that would allow all components of the model to be tested. For this reason, a facilitated crowdsourcing application was developed to collect relevant data and test the model. This application, known as “Building Our Footprints”, was a collaborative project between Land Information New Zealand, Environment Canterbury and the University of Canterbury, and utilised a pool of volunteers to crowdsource a building footprint dataset for the Canterbury region. The project was implemented as a competition, and underpinned by the VGTrust model for scoring, and was targeted towards secondary school students as participants, or the “crowd”.

“Building Our Footprints” was a specific implementation of the model proposed herein, and as such required case specific calibration for the capture of a building footprint dataset. These specific adjustments to the model are discussed in Sections 3.4 – 3.7, with specific detail and background to the application discussed in depth in Chapter 4. Results depicting the performance of the VGTrust model are outlined in Chapter 5. This chapter will focus on how the concepts identified in the literature have been quantified and the VGTrust model developed. It will cover the general algorithmic formula, and its individual components, as well as the implementation of VGTrust using the Python programming language.

3.1 Formulating a Generic Trust Model – The General Elements

One of the primary drivers of this research is to establish a trust rating algorithm that is easily interpreted and can inform any re-user of these data as to the inherent quality of a given feature. To achieve this, the final output of VGTrust - its “trust rating” - will be a measure between 0 and 10, with 0 marking a feature that cannot be relied upon in any way, and 10 being a feature that can be considered of exceptional quality. By establishing this rating system, quality percentiles can be

inferred. Degrees of trust can also be established, and applied to fitness for purpose assessments for data reliance. Both a geospatial professional and a lay-user of VGI could use this proposed trust rating to inform their reliance on a particular piece of VGI, as this rating is the result of a particular weighting of the user, spatial and temporal components of trust.

3.1.1 *The Data Author*

The reputation and expertise of the data author carries significant weight in this algorithm. It is proposed that an author trust rating will provide 30% of the total trust information on a given feature, and be dependent on that data author's perceived expertise with regard to the contributed feature. This weighting is significant, as author trust affects not only spatial quality but also the attribute correctness of CGI. Expertise can be further subdivided into three distinct categories, each of which will require the use and retention of certain information about a feature's creator that is held in its data structure. The first of these is the author's qualification to be operating in the spatial sciences. This could either be as a formal academic qualification, or some form of relevant experience that can be extrapolated to infer a quality level of the particular piece of VGI. In a practical sense, this information could be gathered from a contributor's qualification, or inferred from email domain, i.e. if an email address has an academic or government domain, then the inference is that any resultant CGI contributions could be subject to greater knowledge and subjected to a lesser degree of potential bias or error. The assumptions inherent in such an approach do signal a level of risk, however. In large government organisations, even mapping or spatial organisations, there are many staff who operate in a support or administrative function that may possess very little spatial ability. For this reason a more specific measure of expertise is proposed later in this chapter.

Experience is the second quality concept associated with a data author. It has its roots in social network theory (Golbeck et al. 2008). In essence, a particular network graph, or in this case 'crowd' will enact its own quality assurance through a process of peer ratings, reviews and endorsements. In a facilitated crowdsourcing application, peer endorsement of a spatial feature can be found through the edit history of a feature. In many cases, if a feature has been viewed by members of the crowd but not modified, then this can be considered an endorsement of its quality. Alternatively, experience could be assessed using simple measures such as age, or length of time operating in a given role or field.

The final component of author expertise is what Goodchild (2008) has described as a contributor's "Activity Space". This concept argues that local knowledge surpasses any formal qualification for the contribution of quality CG features. If a feature is contributed by a local about their local environment, then it can be expected that this feature could be given a quality rating that would equal or surpass that ascribed to a contributor with formal geographic training or experience. This is, of course, a qualified example, as only CG features contributed within an author's "activity space" can be deemed to possess this higher assumption of quality.

This study proposes the use of all three measures of author quality – qualification, experience, and activity space. In terms of the latter, the algorithm will calculate trust on the basis that the quality of a feature is inversely proportional to the distance from itself to its author’s activity space. In total, author elements were proposed as contributing 30% of VGTrust.

3.1.2 The Spatial Components

The spatial characteristics of a VG feature are perhaps the most difficult to incorporate into a generic trust algorithm. The context in which data is collected plays a principal role in determining how trust is measured, requiring different approaches for digitised or field based applications. The use of base imagery is a factor for digitising applications, where different imagery products could not only be subject to different map projections, but could also include orthorectification errors in one or more directions. For example, in many places around the world, Google Maps imagery is subject to a 15 metre discrepancy when compared to Google Earth imagery, which leads to a difference in position for any crowdsourced features based upon these respective base layers (Goodchild 2009). In a facilitated crowdsourcing context it is therefore important to use the same base imagery. A field based solution would therefore require an assessment of device precision, or margin of error, information that is available through a device’s metadata. This study proposes a generic VGTrust model for facilitated crowdsourcing applications based on digitising.

To develop a generic VGTrust model, this study used the proxy spatial quality measure of capture scale (de Longueville et al. 2009). The presumption is that if a contributor captures their feature at a closer zoom level, implicit is a greater familiarity with the real world entity, and greater desire to capture additional detail to a greater level of precision. Further spatial measures are proposed for the unique implementation of the model, using building footprints, however these are case specific to each data type. Given the uniquely spatial character of VGI, it is proposed that this will contribute 60% to the final trust calculation. To account for the importance of geometric correctness and precision to CGI re-use, this component has been weighted higher than both the author and temporal components.

3.1.3 The Temporal Components

Temporal characteristics of a crowdsourced feature are often very difficult to quantify. CGI is considered to possess the most merit as a way to measure dynamic phenomena in a given environment, or entities that are subject to regular change. In these cases, the most recently captured CG feature could be considered the most accurate, although this principle does not apply to real world entities that are temporally static, such as an historic church or archaeological site of significance. Nor does this somewhat simplistic approach take into account the principles of the change and contribution ratios proposed by Aragón et al. (2009), or the idea of Crowd

Quality or the Many Eyes principle. For ease of implementation, this study focused on Linus' Law as a means to measure the temporal quality of the data.

Linus' Law focuses on the number of edits made to any given feature, inferring quality through refinement at an exponential scale. The Linus' Law calculation is described in detail in Section 3.3.3.

3.1.4 The Total Model

Whether CGI is fit for purpose should be assessed on a case by case basis. Some information, such as that needed for natural disaster response, is required in near real time, while other situations, such as information on tourist hotspots, are unlikely to require such immediacy. Because of these different but equally valid demands for CGI, temporal quality considerations may be less important for general re-use, and therefore occupy only 10% of the overall trust algorithm.

The overall algorithm can be described as follows:

$$VGTrust = \left(0.30 \left(AuthorExpertise + AuthorExperience + \frac{1}{DistanceToActSpce} \right) + (0.60 \times Spatial) + (0.10(TemporalComponents)) \right)$$

Where Author, Spatial and Temporal Components are normalised to produce results in the range 0 – 10, and are subsequently weighted to produce an overall trust rating between 0 and 10.

Furthermore, it is proposed that the information used will be available as a part of a feature's fundamental data structure, rather than as metadata which for CGI lacks appropriate schema or adherence to applicable standards. The primary information collected and held against a feature will therefore be:

- Author distance to activity space
- Zoom level captured
- Original capture date/time
- Edit history of the feature, including timestamps for each edit

The information calculated and returned for inclusion in the data structure will be:

- Author trust rating
- Spatial trust rating
- Temporal trust rating
- Overall trust rating

3.2 VGTrust.py – Development of Code Package

For this study the VGTrust model was developed into a library of Python functions, collectively known as VGTrust.py. This module in turn was broadly grouped into the three primary components of the model as described above, and included sub-functions to make up each of the author, spatial, and temporal classes. All of these functions were contributed to the overarching VGTrust model function, which drew together all of the parameters to produce a final trust rating. The weightings ascribed to each component of the model were initially trialled at the original modelled values, however these were refined following assessment of the results of the “Building Our Footprints” case study. An extract of these Python functions is depicted in Figure 3.1. A full copy of the Python VGTrust library is contained in Appendix 1 of this document.

The majority of code development was undertaken in Python 2.7.3, using the Wing IDE 101 5.0 interface as a code development and testing tool. These functions were later translated into SQL and Javascript as appropriate, as determined by the needs of the case study application. A number of functions were generic in nature, while some, which accounted for the expertise and reputation characteristics of the data author, were tailored specifically for this case study.

```
# Function to assess the distance from a feature to its creators activity space, using
# northings and eastings

def ActivitySpaceNE(featureX, featureY, creatorX, creatorY):
    import math

    distance = sqrt( (creatorX - featureX)**2 + (creatorY - featureY)**2 )

    if distance <= 1:
        return 10
    elif distance > 1 and distance <= 2:
        return 9
    elif distance > 2 and distance <= 5:
        return 8
    elif distance > 5 and distance <= 10:
        return 7
    elif distance > 10 and distance <= 20:
        return 6
    elif distance > 20 and distance <= 30:
        return 5
    elif distance > 30 and distance <= 50:
        return 4
    elif distance > 50 and distance <= 100:
        return 3
    elif distance > 100 and distance <= 200:
        return 2
    else return 1
```

Figure 3.1 – VGTrust.py model. Extract from Python code module encompassing generic activity space calculation.

3.3 Individual Model Components – Calibrating Case Specific Parameters

A targeted case study was developed to test the proposed VGTrust model, to ensure that all of the required inputs were collected and tested. The “Building Our Footprints” case study was a facilitated mapping application for secondary school students. Students were required to digitise building footprints from aerial imagery, and received a VGTrust score. The generic VGTrust model was calibrated to reflect the specific data collection enabled by the case study. This involved not only the geographic calibration to a scale of the area (in this case a city), but also calibrating expertise and experience calculations on the basis of student participants. Furthermore, a number of additional case specific parameters were included to further refine and customise the model’s veracity and effectiveness. The “Building Our Footprints” application, while running, included a number of these additional parameters that were assessed and built as separate modules, and would specifically apply to the capture of a building footprint dataset. The development of this case study is discussed in detail in Chapter 4.

The general model proposed for assessing trust in crowdsourced geographic information, irrespective of feature type, can be described as:

$$VGTrust = \frac{\sum(Author + Spatial + Temporal)}{10}$$

Where

$$Author = \frac{\sum 0.25(Expertise) + 0.25(Experience) + 0.5(ActivitySpace)}{10}$$

And

$$Temporal = \frac{Linus' Law}{10}$$

The spatial components of the model must be assessed as case specific for each type of geographic feature under investigation. Each feature type, and each physical entity that it represents, exhibits unique spatial behaviours and relationships which must be assessed through incorporation into specific rules. For example, a gravel track could not logically extend over a body of water, or a marina is not appropriately located atop a mountain.

VGTrust proposes a baseline to measure trust for all crowdsourced geographic features, and will illustrate the case specific spatial functions which can be incorporated through the “Building Our Footprints” case study. Several business rules unique to the capture of two dimensional building polygons were modelled in addition to generic trust. Each of the modelled components are described in turn in Sections 3.3.1 – 3.3.3, followed by a summation of these in Section 3.3.4. These sections describe how the generic VGTrust model was calibrated for a specific facilitated collection of building footprint data.

3.3.1 Calibrating Author Trust

Expertise

The attributes contributing to trust in a feature's author can be described in three parts – experience, expertise, and Activity Space. The expertise component of VGTrust required specific calibration for inclusion in a study limited to secondary school participants.

Expertise was assessed as a separate component of this application, and differed from assessment of qualification in its most generic sense in the wider community. A number of studies (Shea et al. 2001, Wai et al. 2009, Bartoschek & Keßler 2013) have been undertaken that seek to associate the spatial ability of secondary school students with a particular subject in the school curriculum, or even a student's preferred subject from that curriculum. Shea et al. (2001) identify a strong spatial ability in secondary school students with an aptitude for mathematics, computer science and physical and natural sciences. Conversely a negative relationship was identified between spatial ability and subjects such as English, social sciences, and humanities. No relationship was perceived to exist between spatial ability and physical education, economics / business or vocational subjects. These conclusions were supported by Wai et al. (2009) who tracked spatial ability in studies ranging from schooling and tertiary education through to vocation. Bartoschek & Keßler (2013) take this concept further by ranking subject associations with spatial ability, which further supports a strong association of this ability with mathematics and physical science, followed by geography, and to a lesser extent humanities and social sciences.

By classifying and ranking subjects according to their association with spatial ability, the VGTrust model was calibrated to deliver an expertise rating accordingly. It was proposed that this component would make up 35% of the total weighting for data author. This factor was also complimented by an assessment of the experience of each participant.

Experience

In the context of secondary schools, experience is largely driven by school year, with the assumption that the greater the number of years spent in secondary education, the greater the exposure of that participant to higher spatial thinking and a greater number of complex concepts. The VGTrust model was calibrated to reflect the age and maturity variation in contributors, with the majority of participants ranging between twelve and eighteen years of age. These age categories were largely reflected by school year, in New Zealand such school grade levels are labelled "Year 7" through to "Year 13" (reflecting the fact that a child begins school at five years of age in school "Year 1"). Trust metrics increased with age, or year at school, with a student in Year 13 of their school education receiving the highest rating for this characteristic. The model was further calibrated to ensure that all possible levels were considered, as a number of secondary schools also include Years generally associated with intermediate schooling – Years 7 and 8. The participation of teaching staff was also acknowledged, with staff receiving a top rating for this

component equal to students in the highest year, Year 13. Experience was proposed as occupying 15% of the data author component of VGTrust.

The resulting model for these two factors – expertise and experience - can be described as:

$$Reputation = \sum (0.35(Expertise) + 0.15(Experience))$$

In the case of “Building Our Footprints”, *Expertise* is representative of preferred subject from the secondary school curriculum, and *Experience* reflects the student’s year at school. *Expertise* has been weighted as contributing 35% towards the overall trust calculation of a feature’s author, with *Experience* weighted lower in the model. This reflects the situation where an ‘expert’ in a given field may, although less experienced than an older contemporary, be more capable of certain tasks within that field, but does not discount the role of experience altogether, as an important aspect of the model.

These two components contributed in total 50% of the value ascribed to the data author in the VGTrust model.

Activity Space

An “Activity Space” (see Section 3.1.1) is used to describe the geographic areas in which a person spends the majority of their time, and is particularly familiar with. This familiarity results in a comprehensive knowledge of that particular area, and reflects the value of local knowledge that is at the heart of all CGI or VGI projects (Goodchild 2007). An activity space, according to Goodchild, could include work, home or frequent recreation areas. Extending this concept, a contributor will have a greater familiarity, and thus precision, for features created within or near to their activity spaces, and that there may be an inversely proportional relationship between the distance of a contributed feature from the person’s activity space, and the quality of that feature. In this case, quality can refer to both the spatial precision of that feature, or, and arguably most importantly, the attribute correctness of that feature (Coleman 2013). Coleman confirms that the most valuable information that comes from local knowledge of VGI contributors is the attribute information that cannot be inferred from interpretation of remotely sensed imagery.

With these activity spaces in mind, and in the case of “Building Our Footprints”, information about two activity spaces were collected for each contributor - the area of the person’s school, and the area in which that person lived. School information was necessarily collected for the allocation of prizes, while home location of participants was only collected at an aggregated suburb level to maintain the privacy of contributors. With these two activity spaces in mind, this aspect of the author component of the model can be described as:

$$ActivitySpaceTrust = \sum (0.5(TrustHomeLocation) + 0.5(TrustSchoolLocation))$$

Where

$$TrustHomeLocation = \frac{1}{FeatureLocation - HomeLocation}$$

And

$$TrustSchoolLocation = \frac{1}{FeatureLocation - SchoolLocation}$$

These two *Activity Space* components contributed 50% of the value ascribed to the data author in the VGTrust model, and were based upon a distance calculation between a crowdsourced feature's centroid, and each respective activity space.

Total Data Author Component

In the case of "Building Our Footprints", these three components of the data author were combined to form one overall calculation of the trustworthiness of that person's digitised features. This model component can be represented as

$$AuthorTrust = \sum (0.5(Reputation) + 0.5(ActivitySpace))$$

3.3.2 Calibrating Spatial Trust

The proposed VGTrust model does not specify particular measures for its spatial component, in recognition of the fact that these will be inherently varied depending on the type of data under assessment. VGTrust does however note that a distinction must be drawn between digitising applications and data collected in the field. In the case of "Building Our Footprints", the model was calibrated for the former scenario. This allows the model to remain sufficiently generic for wide application, while allowing case specific rules to be developed depending on an individual application. For building footprint polygons, this allowed sufficient flexibility to develop and deploy two case specific business rules to augment the base model.

Building footprints are of a specific geometry type, which generally follows a generic set of construction parameters. For example, the majority of buildings when constructed, contain internal angle measurements at their corners of 90° or 270°. Some more modern houses include more gentle angles such as 135°, although the majority of these internal angles remain close to right angles. The first additional trust metric measured each internal angle within a building polygon, classified them based on pre-defined logic, and then issued a base ten rating for each classification to be used in the overall trust model. Given the sometimes imprecise nature of heads-up digitising, these parameters were classified by a deliberately fuzzy schema. A top rating of 10 would therefore be given to any angle that falls within the bearing range 85° - 95°, as an estimate of right angles. A descending level of trust would then be placed in other angles, as depicted in Figure 3.3. An angle calculated as less than 45° would similarly receive a very low rating, as it is unlikely that such an angle would exist as a part of a building footprint. The angle calculation

function is contained as Python code in Appendix 1, and Figure 3.2 depicts the overall process flow of this function.

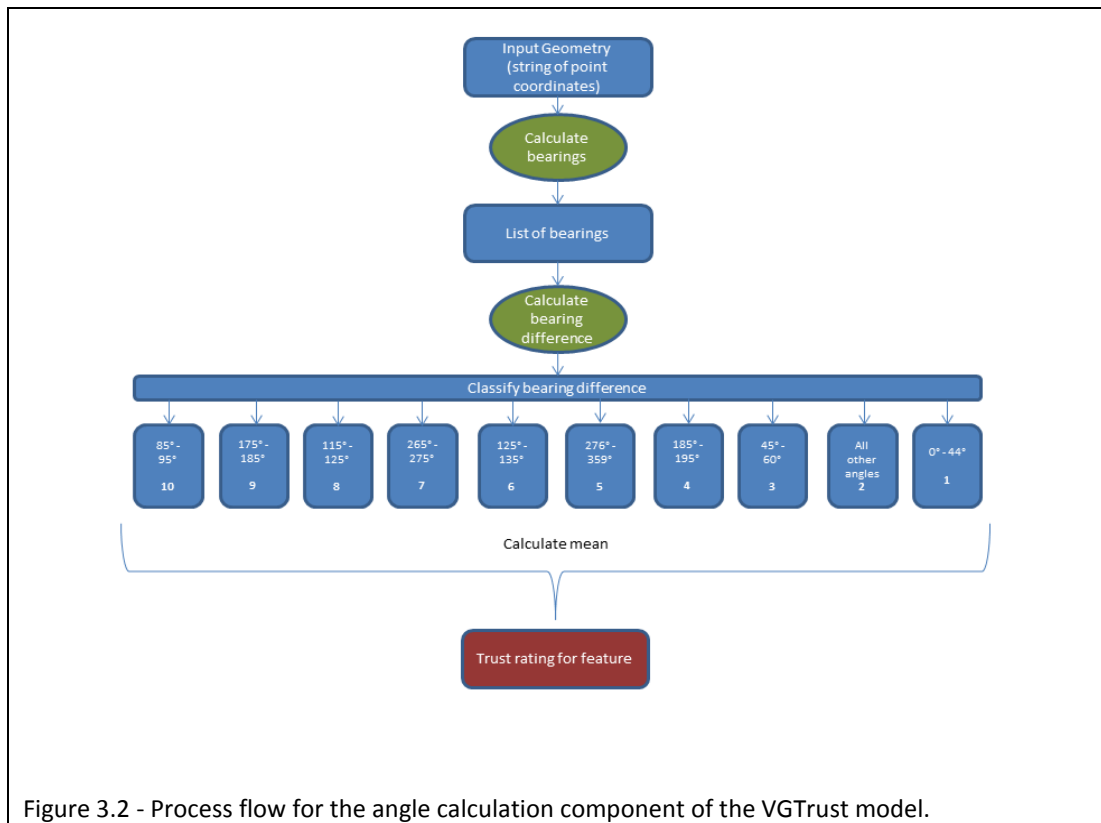


Figure 3.2 - Process flow for the angle calculation component of the VGTrust model.

The second case-specific function involved the classification and rating of the number of vertices for each footprint. This function was based on the assumption that the quality of a footprint can be inferred by the number of vertices it contains. In a digitised environment, a greater number of vertices may indicate a greater level of care and precision taken during the digitising process. A building footprint with fewer than four vertices would logically be a poor representation of its real world counterpart, and would accordingly be rated poorly.

The most important measure of spatial quality in an application based on digitising, is the scale, or zoom level, at which any given feature is created. With the participant determining the extent of every feature from a tiled aerial imagery dataset, the quality and precision of each object in that imagery increases with every incremental step in zoom. When viewed at its smallest possible scale, a building can be viewed with much greater detail, and as such a trust rating awarded based on map scale would be highest at this level. Given the limitations of human perceptions, and the restrictions imposed by an imagery tiling service, the relationship between map scale and feature quality is not linear, but exponential. Therefore in this application, each feature digitised at the smallest possible scale received the highest possible trust rating, with ratings rapidly descending with each subsequent and larger scale step. For example, a feature digitised at the maximum scale of 1:250 would receive a trust rating of 10, however the next scale step out of

1:500, would receive a trust rating of 8. Further at a scale of 1:1000, the next step, a feature would only receive a rating of 5.

“Building Our Footprints” was a crowdsourcing project based on the ability of participants to digitise features on a screen from aerial imagery. For this reason, capture scale as a trust measure occupied a significant proportion of the VGTrust model. Capture scale was combined with the vertex assessment and internal angle calculation to give a measure for spatial trust, specifically calibrated for “Building Our Footprints”, as follows:

$$SpatialTrust = \sum ((Z \times 0.6) + (V \times 0.2) + (A \times 0.2))$$

Where Z = the zoom level or capture scale rating

V = a rating based on vertex count

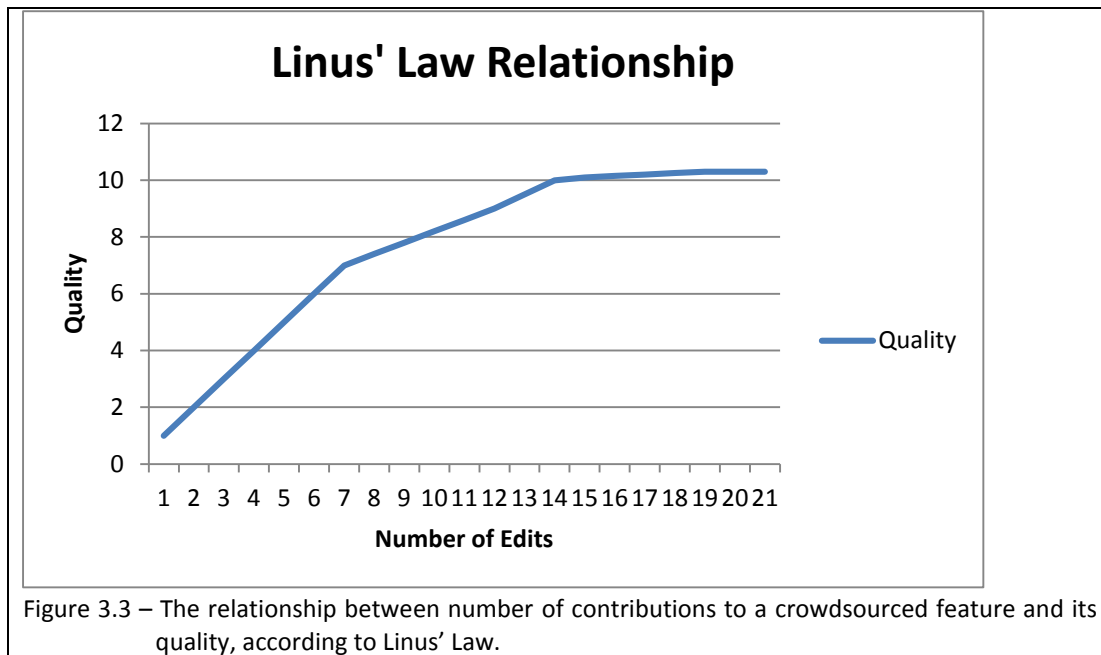
A = a rating based on internal angle assessment

There is interdependency between all three elements in this calculation. Acceptance testing prior to deploying the “Building Our Footprints” application revealed that the capture scale measure needed to be dependent on a simple Boolean rule, reflecting the logic of building footprint geometry. Given that the rating for scale provides a significant component of the overall result, any feature digitised at the maximum possible scale would receive full points for this component, irrespective of whether this feature was in fact a good representation of a building footprint, as defined by the vertex and angle calculations. This could mean that a small triangle, digitised at a scale of 1:250, could receive a trust rating of at least 60%, when it is clear that such a rating is not deserved. The scale rating was therefore modified to only apply when the other spatial functions were satisfied, so that if a feature had a vertex count of 3 or less (i.e. a triangle), it would subsequently receive a rating of 0 for its scale component. Equally, if a feature had a high vertex count but was made up of inappropriate angles, therefore receiving a low rating for the angle calculation component, then the rating for scale could never be more than the rating received for angle assessment.

3.3.3 Calibrating Temporal Trust

There are a number of ways to measure and assess the provenance of a crowdsourced geographic feature. For the “Building Our Footprints” application, Linus’ Law was used as the principal measure of temporal trust. In essence, Linus’ Law describes the effect of multiple contributions to a single feature. Linus’ Law was used for the “Building Our Footprints” application, but did not require any case specific calibration. Irrespective of dataset or data type, the notion remains unchanged that the greatest amount of improvement to a feature is made in its first six edits, with some further, but more minor improvement occurring between six and thirteen edits. When the number of edits exceeds thirteen, the improvement to the feature is negligible. In this way the relationship between edits and quality is

not linear. This relationship was developed into a function based on the relationship depicted in Figure 3.3.



A further function was developed that calculated the change ration of a feature, or in essence its improvement over time, rather than by increased contribution. A vital difference in this case was the concept of endorsement by omission – if a feature remained unchanged by other participants for an extended length of time in an environment with normal contributor activity, then it can be inferred that by not changing a feature, other members of the crowd are endorsing its quality. This function was developed using base ten logarithms when calculating both time difference and number of edits. By using base ten logarithms the effect of large time differences between edits was smoothed. In practice, however, the function did not perform in a way that allowed for reliable use in the model, as the base ten logarithmic calculation returned error values when only one edit was made to a feature. This led to an inconsistent scoring of features in “Building Our Footprints”, which raised practical and ethical concerns due to the fact that the application was a competition. For this reason, the change ration calculation was removed from the final algorithm that was case tested through “Building Our Footprints”. An assessment of this effect presents an opportunity for further research, as discussed in Section 7.1. The Change Ratio calculation is described as follows:

$$ChangeRatio = \frac{\log_{10} E}{\log_{10} D}$$

Where E = number of edits

D = Time difference between first and last edit

The temporal aspect of the VGTrust model was limited to the use of the single Linus' law calculation, although the collection of Change Ratio information was collected as a part of the "Building Our Footprints" competition.

$$TemporalTrust = LinusLaw$$

3.3.4 Full VGTrust Model – "Building Our Footprints" Application

Spatial information is unique as a form of crowdsourced data, in that it has aspects of its quality assessment that do not apply to standard data types. If a spatial data feature is correct in attributes, but incorrect in its spatial definition, then such a feature would be inherently unreliable for its intended purpose. The converse would be true for a non-spatial piece of crowdsourced data, such as a Wikipedia entry (www.wikipedia.com), where it is the text, or attribute quality that is of primary importance. Consequently the VGTrust model ascribes a significant proportion of its total input value to measures of spatial trust, such have been described in Section 3.3.2. The model additionally recognises the importance of other characteristics of CGI, and therefore ascribes the remaining factor weightings to author and temporal components.

When proposing VGTrust, spatial components of trust were assessed as contributing sixty percent of the total. The rating for the author trust was weighted as thirty percent of the total, with temporal trust, in this case restricted to a Linus' Law calculation, forming the remaining ten percent of the calculation, as follows:

$$VGTrust = \sum \frac{((Ax0.30) + (Sx0.60) + (Tx0.10))}{10}$$

Where A= Author trust

S= Spatial trust

T= Temporal trust

For the purpose of clarity, the formula can therefore be expanded as:

$$VGTrust = \sum \frac{\left((0.30(Hx0.25) + (Sx0.25) + (Qx0.35) + (Yx0.15)) \right) + (0.60(Zx0.60) + (Vx0.20) + (Ax0.20)) + (0.10(T))}{10}$$

Where H = Trust Home Location

S = trust School Location

Q = Qualification (favourite school subject)

Y = Year at school (age)

Z = Zoom or scale captured

V = Vertex rating

A = Angle rating; and
T = Temporal rating (Linus' Law)

This form of VGTrust was taken forward for implementation in the “Building Our Footprints” application, the technical implementation of which is described in detail in Chapter 4.

4: “Building Our Footprints” – A Facilitated Case Study

4.1 Background

“Building Our Footprints” is a facilitated crowdsourcing application developed to test the VGTrust model through a specific implementation. Chapter 3 discussed the development of a generic VGTrust model, for use across a wide range of applications, and also addressed how the model was calibrated for a case specific deployment through “Building Our Footprints”. This chapter will discuss how this application was implemented, including database design, an engagement model targeting specific participants, and the deployment of VGTrust.

The key driver for “Building Our Footprints” in the context of this research was to demonstrate how the VGTrust model could be successfully deployed to facilitate geographic data collection by government agencies. It also allowed for the specific collection of all relevant inputs required by the model, and meant that each component could be thoroughly tested and analysed for relevance, before finalising VGTrust. “Building Our Footprints” was therefore developed with the support of Land Information New Zealand (LINZ) and Environment Canterbury (ECAN), as a part of the Canterbury Spatial Data Infrastructure (SDI) Programme. The Canterbury SDI Programme sought to establish a local SDI for geographic data, in order to generate cost savings and improve the efficiency of the region’s post-earthquake infrastructure rebuild.

Within the Canterbury SDI Programme, a more effective means of managing property information was proposed, which relied on a number of property specific datasets. This work stream was labelled the “Property Data Management Framework” (PDMF), and sought to improve consistency in the way that government agencies, local government agencies, and utilities companies, managed property information. A building footprint dataset was identified as a principal component for the management of property information, both for emergency response, and for ongoing information management. The onus for the capture and management of these datasets lies with individual territorial authorities, responsible for their own districts. The nature of the Canterbury region means that there are multiple territorial authorities within the area that do not possess a comprehensive building footprint dataset. Specifically, the Waimakariri and Selwyn Districts did not possess these data, and indicated that crowdsourcing was an appealing, low-cost option with which to gather the footprints required. The Waimakariri and Selwyn Districts are those immediately abutting Christchurch City, and include satellite settlements of notable population. The standard way to generate this dataset is through the use of feature extraction software from multi-spectral remotely sensed imagery. An alternative is for larger agencies to dedicate

staff to the digitising process of features from aerial imagery. These options however, are unrealistic for many small or medium sized agencies due to high cost.

Crowdsourcing, therefore, presented an attractive option to meet these needs. Crowdsourcing allows outsourced production to a community of online users for minimal cost. A crowdsourcing application also fits with the ethos of open data and data sharing, as promoted by the Canterbury SDI Programme, and it was determined that this crowdsourcing project was particularly timely, in that it met a specific business need at a crucial junction in the Canterbury SDI work programme.

A further key component of any SDI is the capability, literacy and capacity of all of the actors within the data ecosystem (<http://www.linz.govt.nz/about-linz/our-location-strategy/geospatial-strategy-and-work-programme/what-sdi>). A principal goal of the New Zealand Geospatial Office is to improve the geospatial capability and capacity of New Zealand, by targeting both secondary and tertiary education sectors, and encouraging the progression of new participants into the spatial sciences (<http://www.linz.govt.nz/about-linz/our-location-strategy/geospatial-strategy-and-work-programme/new-zealand-geospatial>). The New Zealand Geospatial Office regularly sponsor “virtual fieldtrips”, run through the provider company LearnNZ (www.learnz.co.nz) targeted at high school communities, to expose students to spatial concepts and career paths. In line with the Canterbury SDI Programme and the PDMF work stream, the theme of the 2014 virtual fieldtrip suite was centred on property information management and building footprints¹.

4.2 Competition Implementation

While a building footprint dataset is essential for the effective management of property information, not all areas in the Canterbury region were in possession of such data. As noted above, the development of a Canterbury SDI provided a crucial test bed for this application. The Waimakariri and Selwyn districts were specifically targeted for data capture using “Building Our Footprints”. The Waimakariri District occupies an area to the North of Christchurch City, and includes the relatively populous settlements of Rangiora, Kaiapoi, and Amberley. The Selwyn District, to the South, includes the rapidly expanding Rolleston Town, Lincoln, Prebbleton, and Rakaia townships. Both districts also include a scattering of smaller rural villages.

Given the regional focus, the most appropriate location to house the infrastructure for the application was on the newly established Canterbury Maps

¹ The crowdsourcing of building footprints in the Canterbury region presented a means for students participating in the virtual fieldtrips to gain “hands-on” experience with geospatial technologies, and participate in a project with clear and measurable benefits to the Canterbury rebuild. Furthermore, there are a number of geospatial datasets, that because of their significant abstraction from the real world features they represent, are not easily understandable for people not involved in the spatial sciences, and in this case, by secondary school students. A building footprint is a concept that is easily understood by people without in depth geospatial training, and therefore ideally suited for a crowdsourcing application.

(www.canterburymaps.govt.nz). Canterbury Maps a regional spatial data viewer hosted by Environment Canterbury (ECAN), and was developed as a part of the ongoing implementation of a Canterbury SDI.

The implementation of “Building Our Footprints” is described in the following two subsections. Section 4.2.1, will explore the social context of deploying a competition to secondary school students, including how participation was incentivised and encouraged. Section 4.2.2 examines the technical design and implementation of the application.

4.2.1 Engagement Model for Participation

ECAN and the New Zealand Geospatial Office have a strong presence in secondary schools across the region; promoting both environmental concerns and the fostering of spatial capability amongst youth. The competition for secondary schools and students received support from both organisations, and was structured as follows.

1. The competition, ran for one month from its commencement on 28 July 2014, and sought contributions from secondary school students. The students registered for the website, and digitised building footprints from a composite of aerial images across the region.
2. For each footprint submitted to the database, the VGTrust model allocated a rating between 1% and 100%, depending on its assessment of predetermined factors (see Chapter 3).
3. These calculated trust ratings informed the scoring for the competition – for each footprint successfully captured with a trust rating in excess of 75%, a point was allocated to both the individual and school.
4. The individuals and schools competed for prizes. Financial support for this project, in the form of these prizes, was gratefully received from LINZ via the Canterbury SDI Programme. The prizes were of a quantity and quality to ensure widespread participation from a variety of school and students, such as an iPad, money, and movie tickets.

The available prizes were a principal motivator, providing a tangible financial reward. An additional motivator, particularly appropriate in the Canterbury region as it recovers from the earthquakes of 2010 and 2011, was what Bartoschek & Keßler (2013) identified as “the belief in the social benefit of their contributions”. The project appealed to the students’ desire to contribute to the rebuild of their region, and promoted participation as both easy and of significant value. The following key message was delivered to potential participants via a range of communications channels, including, classroom delivery, poster advertising, promotion in local and national youth publications (www.tearaway.co.nz), social media, and other online channels. The key message delivered to potential participants was that their contribution, while being easy to make, contributed meaningfully to a range of rebuild functions, and response to possible future natural events. This message focussed on the ability of emergency services to

respond more quickly and with greater precision should our property information be arranged according to its building allocation, or building footprint. This key message was delivered in a way that made it easily comprehensible by the spectrum of ages and abilities present at high schools, in order to emphasise the meaningful influence each contribution will make to a student's own future environment.

"A good building footprint dataset is an essential part of managing property information, and can be used for many different tasks, including emergency services response and disaster recovery.

By participating in this project, students will not only compete for a chance to win prizes for their school and themselves, but will also contribute vital information for our property management, and for our future emergency response.

There are cash prizes for individuals and schools, ten double movie passes to give away, and an iPad mini up for grabs!"

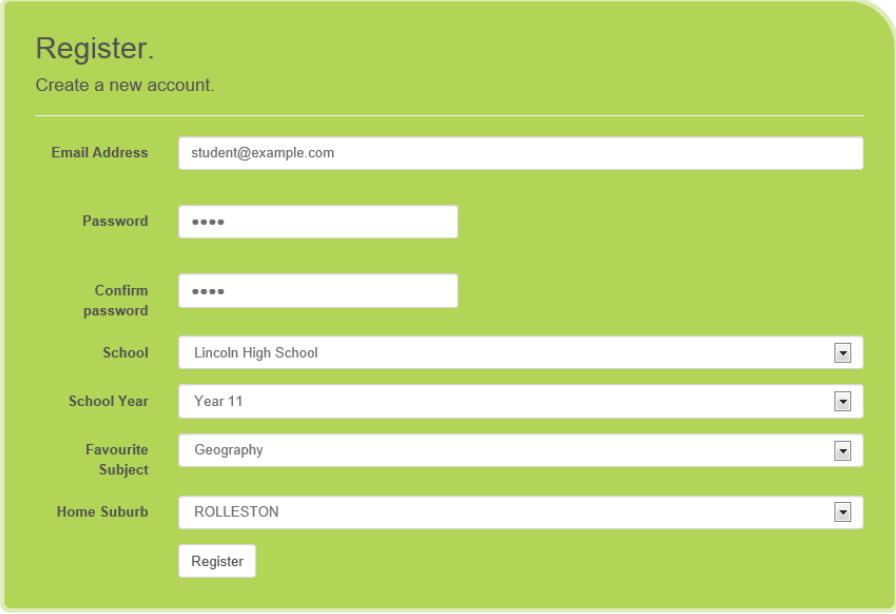
Finally, the project sought to include a third motivating factor, being the desire of students to be instantly gratified with recognition of their contribution, as previously noted by Bartoschek & Keßler (2013). Once a building footprint is committed to the database, it was immediately rendered back onto the map canvas with its authorship and trust score as viewable metadata, as well as being reflected in the colour of the feature. Participants could subsequently re-edit the feature, if its trust rating had not reached the 75% threshold for the award of points, or see their trust rating rendered as an acceptable shade of green, and continue to other areas for further contributions. As each feature, once rated by the VGTrust model, was submitted back onto the map with an associated colour indicator of trust, it was easy for a participant to immediately receive gratification for their contributions. The trust rating generated was coloured on a graduated scale between red (for a low trust rating) through to a deep green (for ratings achieving above the 75% threshold).

4.2.2 "Building Our Footprints" System Architecture

"Building Our Footprints" was implemented as a cross-agency project between LINZ, ECAN, and the University of Canterbury. Hosting of the application was through the Canterbury Maps web portal.

The application was built as an ArcGIS Online application, using ESRI web feature services. Before mapping could commence, users were prompted to register for the site, thereby providing all of the relevant personal details required for accurate assessment by the VGTrust model (an example of this capture is shown in Figure 4.1). The only identifying personal information collected at this stage was the participant's email address, which was used primarily as a tool to recover user passwords should they be lost, and to contact participants for the allocation of prizes. No other identifying personal information collected, as noted above, with information such as the participant's home location being aggregated to suburb

level. All participant data was stored securely behind Environment Canterbury firewalls and proxy servers.



Register.
Create a new account.

Email Address:

Password:

Confirm password:

School:

School Year:

Favourite Subject:

Home Suburb:

Figure 4.1 – The information required to implement the VGTrust model was collected from participants at the time of registration.

Once all registration information was collected, the user was redirected to the map to begin the capture process. Each building footprint was delivered as a part of an editable ESRI feature service layer, and was committed to the database upon completion. The primary database consisted of a table structure including school locations, suburb locations, a restricted school subject domain, and finally a table to capture details of the feature, and all of its associated parameters. The base imagery used for the mapping process was captured, stored and orthorectified into a mosaic by ECAN, which meant that all participants were generating data in relation to the same base imagery, or context data. This removed one possible source of ambiguity within the model, and is consistent with any facilitated CGI application.

During the participation period, it was necessary to trigger the VGTrust model each time a new feature was committed to the database, as well as each time an existing feature was edited. The regularity of this function meant that results were generated in a timely manner, and also allowed for multiple edits to be made to any given feature in near real time. The complex nature of this triggering function necessitated a translation of VGTrust from Python to Structured Query Language (SQL). The full translation of the model is annexed as Appendix 2. Following each trigger and running of the model, an additional, non-editable, layer was rendered onto the map, displaying the trust rating of the footprint feature. This appeared as a colour-classified square at the centre point of each polygon, as was non-editable. The trust rating layer refreshed automatically after one minute, and both the footprint and the trust rating layers were able to be queried with intuitive pop up functionality, as shown in Figure 4.2 below.

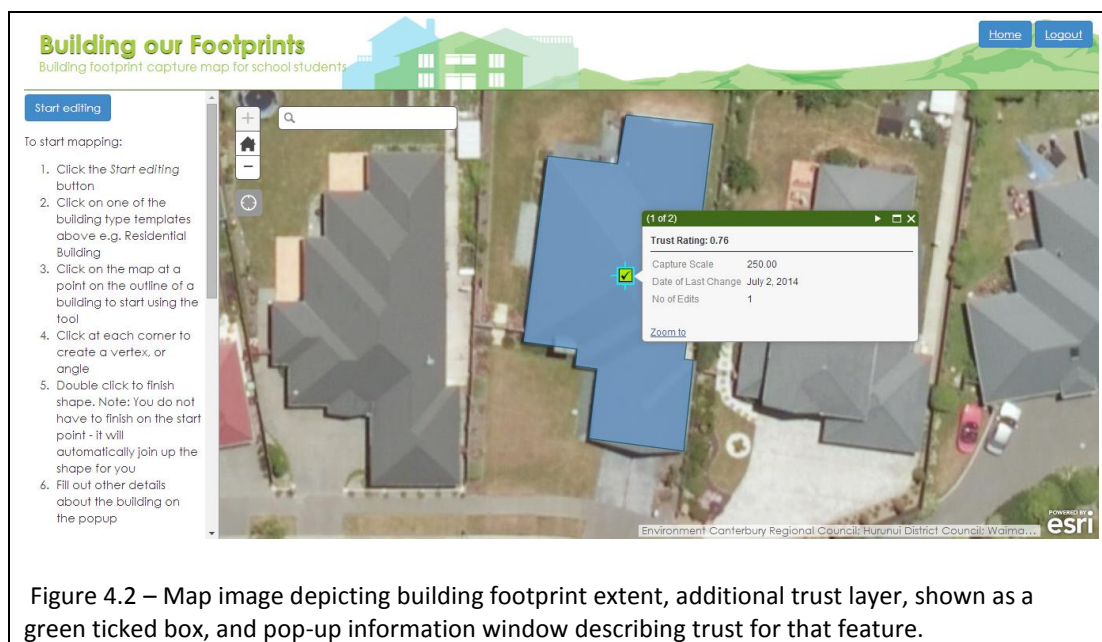


Figure 4.2 – Map image depicting building footprint extent, additional trust layer, shown as a green ticked box, and pop-up information window describing trust for that feature.

The competition ran for one month from 28 July 2014, attracting 42 participants, and resulting in the capture of 18,792 individual building features exceeding the 75% trust threshold. These were subsequently analysed against a set of reference data to determine whether this implementation of the VGTrust model generated statistically and geometrically correct results.

4.3 Quality Assurance Reference Data

In order to assess the performance of the VGTrust model, a comparison was made to data that was considered trustworthy. No existing dataset provided a complete or accurate coverage of building footprints within the Canterbury region. To resolve this lack of reference data, a dataset was digitised by a group of five Masters and PhD students at the University of Canterbury, whose expertise and precision was deemed sufficient for the creation of a comparison dataset. These footprints were captured within a set of ten randomly generated plots across the city of Christchurch and its surrounds, using the “Create Random Point” function in ArcGIS Desktop version 10.2. Each plot measured approximately 31,415 m², the result of ten randomly generated points which were subsequently buffered to create circular plots with a radius of 300 metres around each of the points. Within these areas, building footprints were digitised to a high standard of accuracy for use as a reference dataset, containing a cross section of building types. To ensure consistency, the digitised buildings were created from the same imagery utilised in the “Building Our Footprints” application – sourced as a web mapping service from the Environment Canterbury public facing REST services. All footprints were digitised at the maximum possible scale, and were visually peer reviewed by another member of the group.

This reference dataset was compared with the data collected by student volunteers as a part of the competition, by utilising the ‘intersect’ function contained in ArcGIS 10.2. This process extracted a subset of 530 features which were subsequently analysed.

4.4 Summary

“Building Our Footprints” was designed and implemented to collect all of the input data required for VGTrust. This thesis is the first known research on CGI where many of the trust indicators were combined into one metric. It was therefore necessary to specifically collect these data, and the application achieved that goal.

Furthermore, and arguably most significantly, “Building Our Footprints” signalled an appetite by authoritative mapping agencies to experiment with CGI in their spatial data supply chains. The data collected through “Building Our Footprints” is presented and analysed in Chapter 5.

5: Results

This chapter will present an overview of the data obtained from the “Building Our Footprints” mapping competition, and illustrate the trends and relationships inherent in the components of the dataset. This process will describe the extent and the nature of the data obtained, the result of its comparison to a reference dataset of building footprints, and the relationship of the overall quality assessment to each of its individual components. This will lead to a confirmation of relevance for each component of the VGTrust model, based on whether that component contributed materially to the generation of a reliable quality metric. Finally, these results will be used to confirm the appropriate weightings of each factor in the model. It is noted that the completeness of the building footprint dataset is not under investigation in this section. Rather these results provide a comparison between crowdsourced data and reference data, where both exist to model a particular feature, in this case a building footprint.

5.1 The Building Our Footprints Mapping Competition

The “Building Our Footprints” mapping competition, as described in *Chapter 4 – “Building Our Footprints” – A Facilitated Case Study*, ran for one calendar month from 28 July 2013 to 28 August 2013. During the course of the competition, 42 participants digitally captured a total of 18,792 unique building footprint polygons. There was a diverse range of participants from eight schools across the Canterbury region, including students from all school levels and some staff. The majority of participants had little experience creating maps prior to the competition, although many were familiar with a variety of electronic mapping products, such as the plethora of apps found on phones, computers and other devices. The resultant dataset contained footprints of varying quality - a high number of excellent polygons were countered by a series of poorly created examples. A number of errors were revealed that indicated a lack of familiarity with the digitising tool – ArcGIS Online – predominantly with the finishing of a polygon, and where the most appropriate location was to “double-click” in order to complete capture. As a result, a comprehensive data cleaning process was undertaken to remove any features that, as outliers, could affect the analysis of results.

5.2 Reference Data

A reference dataset was generated specifically for this project, by a group of five students and staff associated with the Masters in GIS programme at the University of Canterbury (www.mgis.ac.nz). The data obtained from these experts was deemed to be of a suitable standard to be trustworthy, due to their past experience and expertise. A selection of these data were randomly peer reviewed as a quality

assurance measure, whereby they were assessed manually for geometric correctness and positional accuracy in terms of the underlying aerial imagery. The process of generating reference data is described in Section 4.3.

This collection of reference data created significant redundancy, as only a relatively small number of features overlapped between the reference dataset and the “Building Our Footprints” dataset. An intersection query run through ArcGIS 10.2 revealed a total of 530 building footprints that correlated between the two datasets. Of these, a number of features obtained from the competition were clear outliers – the result of either severe misunderstanding of the subject matter or of malicious intent. Figure 5.1 depicts an example of student capture where the participant has erroneously captured a school’s sports fields rather than the relevant building infrastructure. Several students also showed a desire to capture the extent of features of interest to them, such as whole parks or schools, instead of a specific class of feature. This motivation is generally associated with more traditional crowdsourced maps, such as Wikimapia, where users are motivated to create points of interest, rather than to generate specific and targeted datasets.



Figure 5.1 - Feature capture at a school site. Participants have captured all features of interest, and have not been limited to buildings. Such features could include sports fields, parks, and recreational infrastructure such as tennis courts and swimming pools.

5.3 Processing of the Raw Crowdsourced Data

Throughout the mapping process, features created by participants were rated according to the weighted VGTrust model proposed in Chapter 3 and Chapter 4, where an arbitrary threshold of 75% trust set for acceptance of a feature. For analysis of results, all data was carried forward, irrespective of the quality rating inferred by the original model. For this reason, a large and diverse range of features were analysed, some of which were inappropriate for further assessment. For example, the polygons representing fields and sports facilities, as depicted in Figure

5.1, were identified through an intersection query and included in the original data extraction. Similarly inappropriate were polygons captured with major geometric errors such as gaps, overlaps, and slivers, which were also included in the original resultant dataset. In order to conduct an accurate analysis of the VGTrust model, these geometric and statistical outliers were removed. This process is discussed further in Section 5.4.

QGIS 2.0 Dufour was used to conduct a simple intersection query between the crowdsourced building footprint dataset and the previously captured reference dataset. This resulted in a sample dataset of genuine building footprint polygons for comparison, containing 453 individual features, and allowed a direct comparison between the reference data and the crowdsourced data. This sample reflected a manual process of identifying erroneous capture, such as non-building features and geometric errors, and removing these from the dataset. The sample was of a sufficient size for comprehensive statistical analysis. Further cleaning was also undertaken during a subsequent phase of statistical investigation, to remove outliers following a regression analysis which revealed high residual values.

5.4 Independent Quality Assessment

5.4.1 Object Quality

To assess the effectiveness of the modelled variables, an independent quality measure for each building was established. Zhan et al. (2005) propose a method of assessing the “Object Quality” (OQ) of a feature when comparing its capture in two different datasets, for building extraction from remotely sensed imagery. The approach is equally valid for crowdsourcing applications. OQ hinges on percentage overlap between the two datasets, both the area of intersection, as well as the areas that fall outside of the overlap. These outside areas could be either those that should be a part of the building and not captured, or areas that have been captured and are not correct. This process is illustrated in Figure 5.2 for a series of buildings, and results in a percentage ratio that is represented by the following equation:

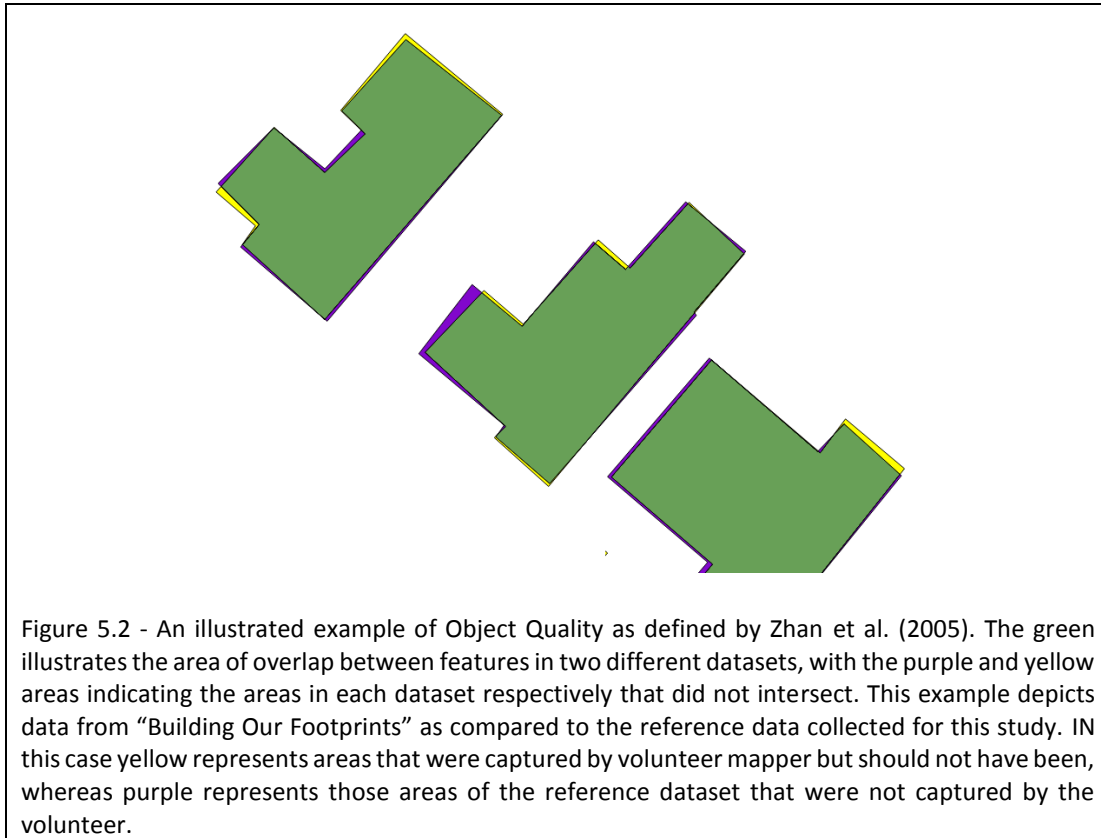
$$OQ = \frac{i}{i + o + u}$$

Where *OQ* = *Object Quality*

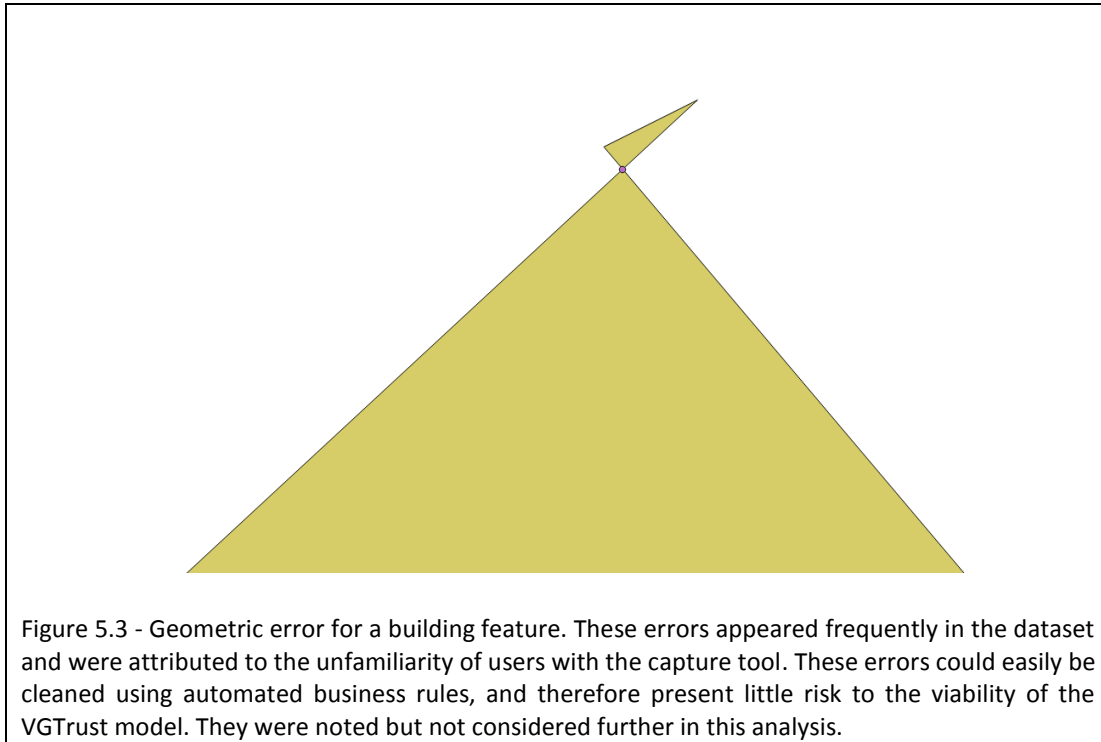
And *i* = *intersection*

And *o* = *area over*

And *u* = *area under*



The OQ analysis was carried out in QGIS using the “Intersect” geoprocessing query, as well as the pre-built “Difference” tool to establish the areas in each dataset outside of the intersection area. These area calculations were then used to establish OQ measures for each footprint in the sample. The dataset was also checked for geometric errors, an example of which is illustrated in Figure 5.3. For this analysis these errors were noted but not included in the Object Quality assessment, as they were largely a by-product of the digitising tools used for capture. The purpose of the model was to facilitate the integration of crowdsourced data into authoritative datasets. As these types of geometric errors could easily be filtered and corrected through automated means in a given application, and were therefore not considered further in this analysis. Instead, the quality assessment in this research focussed on the general nature and extent of captured crowdsourced features.



5.4.2 Statistical Data Cleaning

Following the geometric data cleaning described in Section 5.3, with a subsequent comparison to OQ, the resultant dataset was imported into the Minitab 17 statistical software package for more detailed analysis. The dataset was standardised for both OQ and the previously generated VGTrust ratings. Both of these metrics were also initially analysed in comparison to one another, using basic descriptive statistics (mean, standard deviation) in addition to a linear regression function. These analyses identified a number of statistical outliers showing unusually large residual values. These features were removed from the dataset in order to prevent any undue influence of extreme values on the result. In many cases these were due to very small polygons influencing the area ratio calculations, in particular when polygons contained in the same dataset overlapped one another. This situation is depicted in Figure 5.4, revealing where a sliver polygon could unduly affect the statistical outcome. Following this process, 403 individual building polygons remained in the dataset for further analysis.

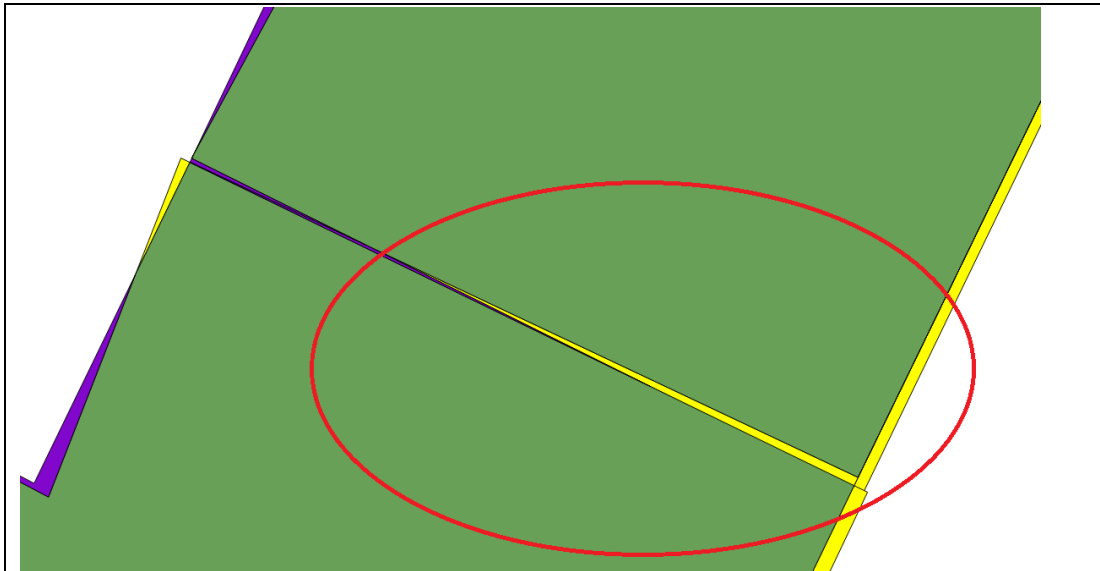


Figure 5.4 - The sliver displayed above is the result of an overlap between two adjacent buildings in the same dataset. The extremely small area meant that the feature generated a high residual value, resulting in its exclusion from the subsequent analysis.

5.5 Analysis of the VGTrust Model

A dataset of 403 individual features was further analysed using the Minitab 17 statistical software package. This dataset included information on the “Object Quality” of those features, the VGTrust values as originally modelled, as well as data pertaining to the majority of inputs used for the original modelling of trust. An initial analysis was undertaken of each of the quality assessments – “Object Quality” and VGTrust - and these values were then compared with each other for any similarity or dissimilarity of statistical significance. “Object Quality” was then compared to each contributing variable in turn, to assess the impact each of these had on the final outcome, using a regression equation. Furthermore, a multivariate analysis was conducted on the contributing variables in order to determine if there was a significant relationship between these. Finally, a response optimised multiple regression was run, to determine an optimum level with which to weight each of the contributing variables. The results described as follows will be presented in this order, after which these will be discussed in Chapter 6, and a final VGTrust model format will be proposed.

5.5.1 Attribute Correctness

The following statistical analysis was undertaken as a measure of **geometric** correctness for crowdsourced features, and did not include a measure for **attribute** correctness. The nature of the data collection meant that attribute capture was optional, and the crowd participants displayed a tendency to omit fields from the data schema that were not mandatory. Anecdotally, a small number of participants supplied some of this information, such as labelling a shop as a “supermarket” or a

“gym”, or in some cases “my house” or “my Mum’s house”. Ultimately a formal quality assessment of feature attributes was not undertaken, and the statistical analysis focussed in its entirety on the geometric correctness of the features. This does not, however, indicate that attribute correctness is less important in a crowdsourced dataset than geometric correctness, rather, it signifies a lack of relevant data to appropriately test the model. These assessments should be the subject of further research, and are necessary in order to establish a holistic model for assessing all aspects of these data. The potential of this topic for future research is discussed in Section 7.1.

It is also important to note that a number of the factors modelled in the VGTrust algorithm are predictors for attribute correctness. For example, in a digitising environment, a contributor’s activity space is less relevant for establishing geometric correctness, but very highly associated with attribute correctness, by leveraging local knowledge that cannot be extracted through the interpretation of aerial or remotely sensed imagery. As the following analysis does not focus on attribute correctness, a number of the components of the model have generated less significant results than originally anticipated. Importantly, this does not mean that these should be discounted from the VGTrust model moving forward, but warrant discussion here and further analysis in subsequent research.

5.5.2 Object Quality vs VGTrust

An initial analysis was undertaken to determine the relationship between the measure of OQ and the originally proposed values for VGTrust. The sample size of 403 features was sufficient to detect trends of significance without needing to normally distribute the data, although a normal distribution was also obtained as a point of comparison.

Using Minitab 17, a 2-Sample t-Test was performed which detected a statistically significant difference between the two means of OQ and VGTrust, of 0.14948, approximately 15% of trust for a feature. A low p value of $p < 0.05$ indicates that the difference between the two means is significant, and this is further reinforced by the relatively tight 95% confidence interval. The results of this test are displayed in Figure 5.5. These results also indicate a notable difference in distribution between the two measures of quality, and this was explored further through a comprehensive analysis of each measure.

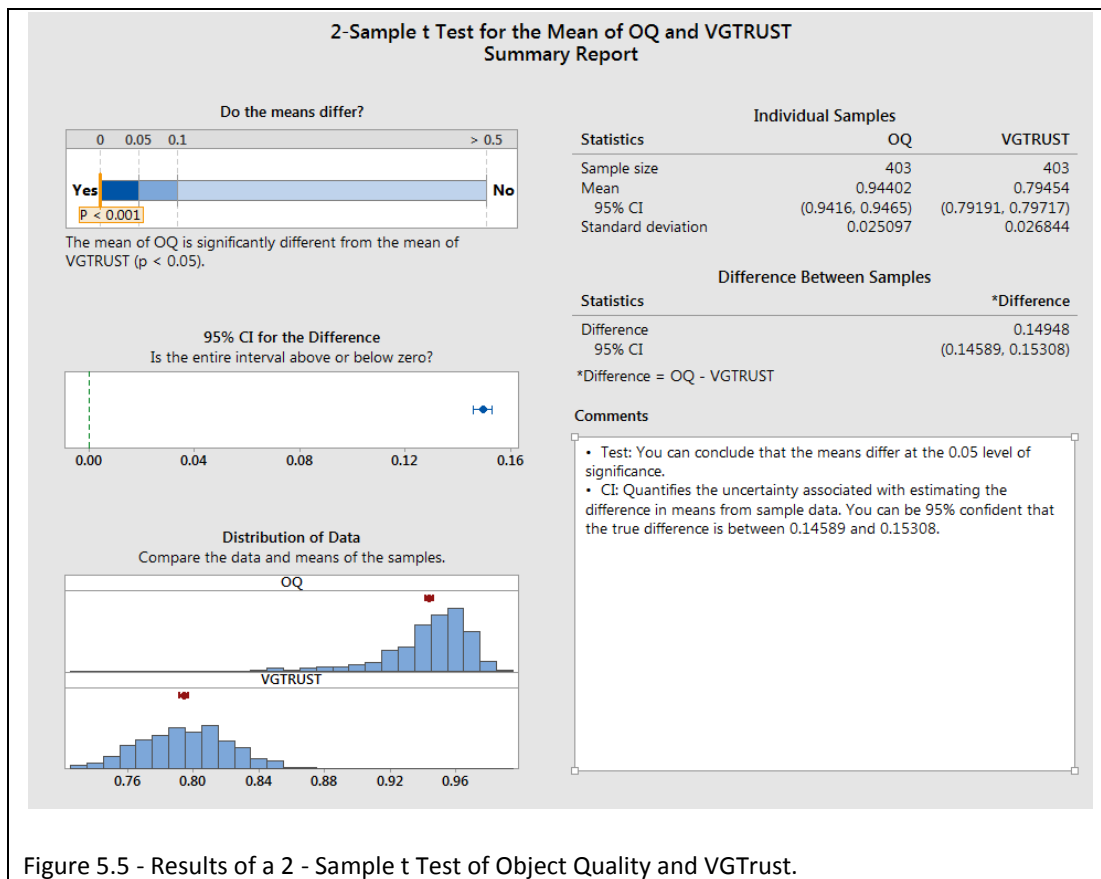


Figure 5.5 - Results of a 2 - Sample t Test of Object Quality and VGTrust.

VGTrust was identified as being inherently normally distributed, but was restricted to a tight range between 0.72733 and 0.87275, with a mean of 0.79454. OQ in contrast revealed a significantly higher overall range of values, between 0.84069 and 0.98611, and with a mean of 0.94402. Interestingly, both OQ and VGTrust samples exhibited similar standard deviation values, with 0.025 and 0.027 respectively. A comparative illustration of these two distributions is depicted in Figure 5.6, showing a normal distribution for VGTrust and a notable left skew distribution of OQ. As previously noted, due to the sample size of 403 individual features, the difference in distribution would not affect the result of the t-Test conducted.

It is important to reaffirm at this point that OQ represents an analysis of geometric quality for each feature in the sample, whereas the original VGTrust model was calibrated to account for attribute correctness also. For this reason, OQ reveals a consistently higher range of values than VGTrust, and the interpretation of results will focus on the geometric qualities of the data when drawing inferences from these results. The result of the 2-Sample t-Test did however signal that there is a statistically significant difference between VGTrust and OQ, which was explored further using a number of regression analyses.

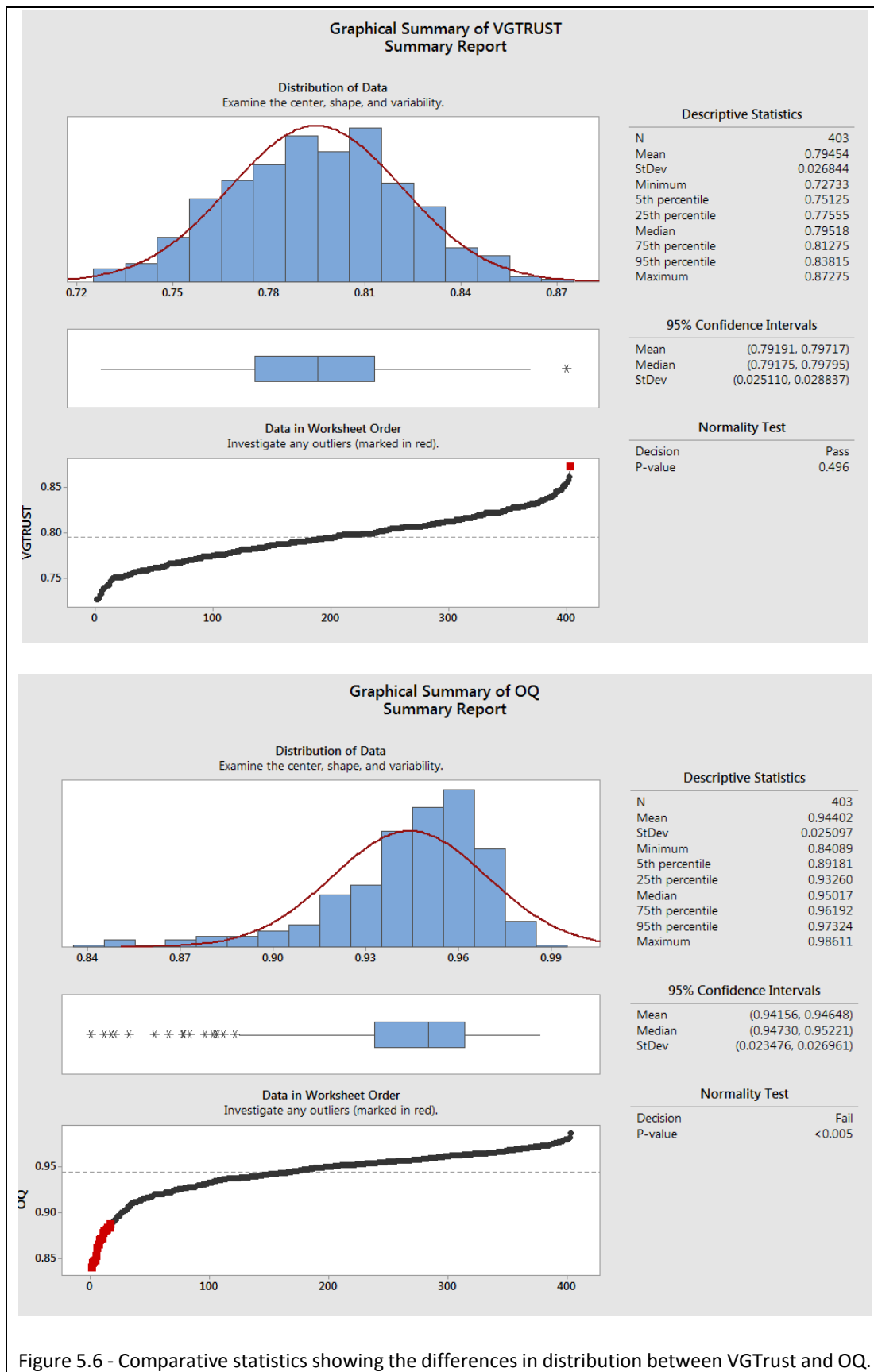


Figure 5.6 - Comparative statistics showing the differences in distribution between VGTrust and OQ.

5.6 Exploring the Modelled Variables

The initial VGTrust model was made up of a number of variables, categorised into three groups – measures of trust for a feature’s author, specific spatial measures, and temporal characteristics relating to the provenance of a crowdsourced feature. Through the “Building Our Footprints” data capture, each of these factors were quantified by using case specific examples, for example a data author’s expertise and experience was based around their year at school and preferred subject respectively, and spatial precision was inferred by the scale at which a specific building feature was captured. These factors were analysed to assess the significance of their contribution to the overall model, through reference to the independently obtained measure of OQ, through a series of regression analyses. A further multiple regression procedure was carried out to optimise the factor levels within the model, and a principal components analysis was undertaken to determine if a statistically significant relationship existed between various combinations of these factors.

The following factors analysed further, and are discussed in turn:

- Expertise – represented by the participant’s preferred subject at school;
- Experience – Expressed as the participant’s year at school;
- Activity space – represented by the distance between the feature and the participant’s home and school;
- Spatial precision – the scale at which a feature was captured, as well as the impact of vertex count;
- Temporal trust – the number of edits made to each feature.

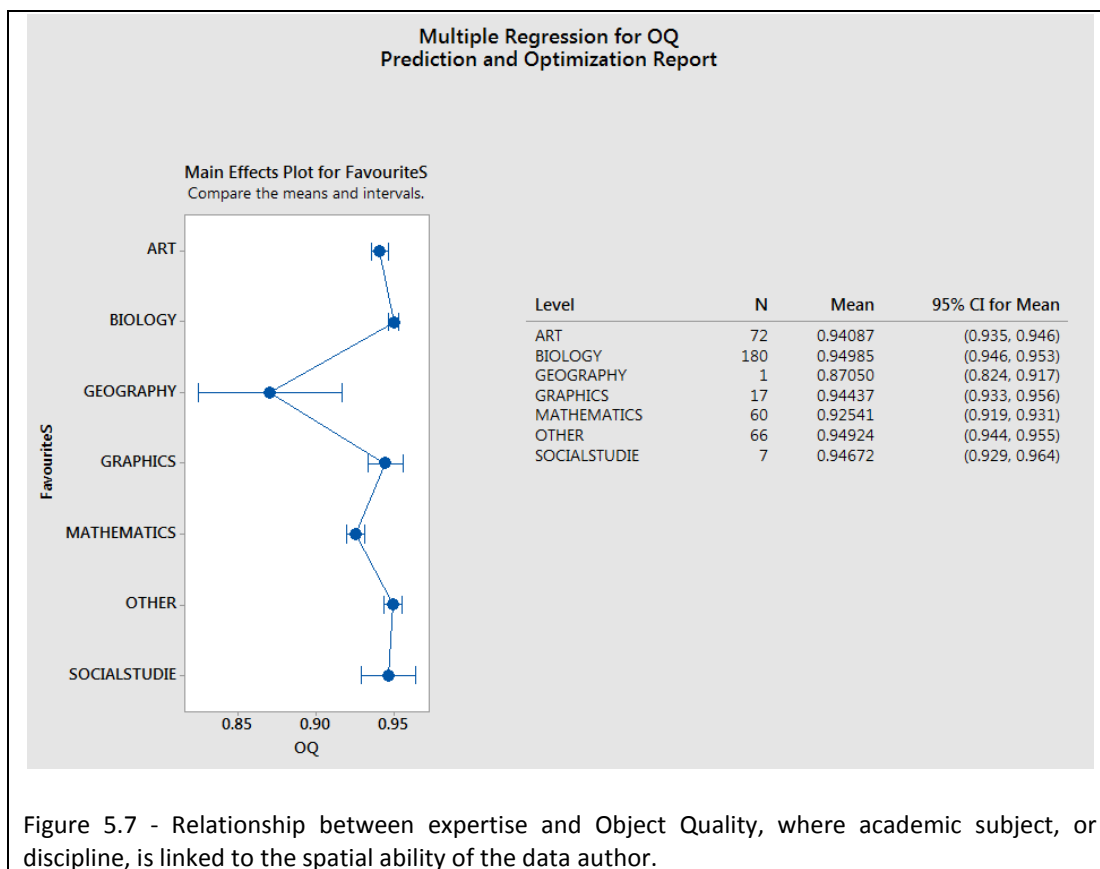
5.6.1 Expertise

Expertise information was captured in this dataset as each participant’s favourite subject at school, and was modelled on the studies on the association between spatial ability and subject expertise (Wai et al. 2009, Shea et al. 2001). These studies proposed a link between certain subject matter expertise and spatial ability, and the results in this study did go some way to reinforcing this hypothesis. Figure 5.7 depicts the relationship between OQ and the preferred subject provided by each contributor. Unfortunately the sample size ($n = 403$) did not provide sufficient data to test the association between some subjects, in particular Geography and Social Studies, the latter including Geography at junior year levels. In addition, 66 features were associated with the subject choice “OTHER”, which unfortunately does not allow an appropriate level of analysis or clear association with spatial ability.

Given the tight clustering of OQ values, the confidence intervals for mean OQ values by subject were tight. In most cases, the results aligned with predictions of spatial ability based on a subject classification, or tiered approach.

This study's results partially endorse the results of previous studies, in particular the highest mean OQ score for Biology (0.94985), with a notably tight 95% confidence interval. This performance of a physical science subject is consistent with the prevailing literature. Mathematics, Graphics and Art also produced high mean OQ values, although the 95% confidence intervals for these subjects were broader than that for Biology. Social Studies presented a high mean but large confidence interval, likely attributable to its small representation in the sample ($n = 7$). Geography showed a lesser association with spatial ability in this sample, with a mean OQ value of 0.87050, which is low for the OQ range in the sample. Unfortunately this conclusion cannot be statistically relied upon, as features associated with geography were under-represented in the random sample.

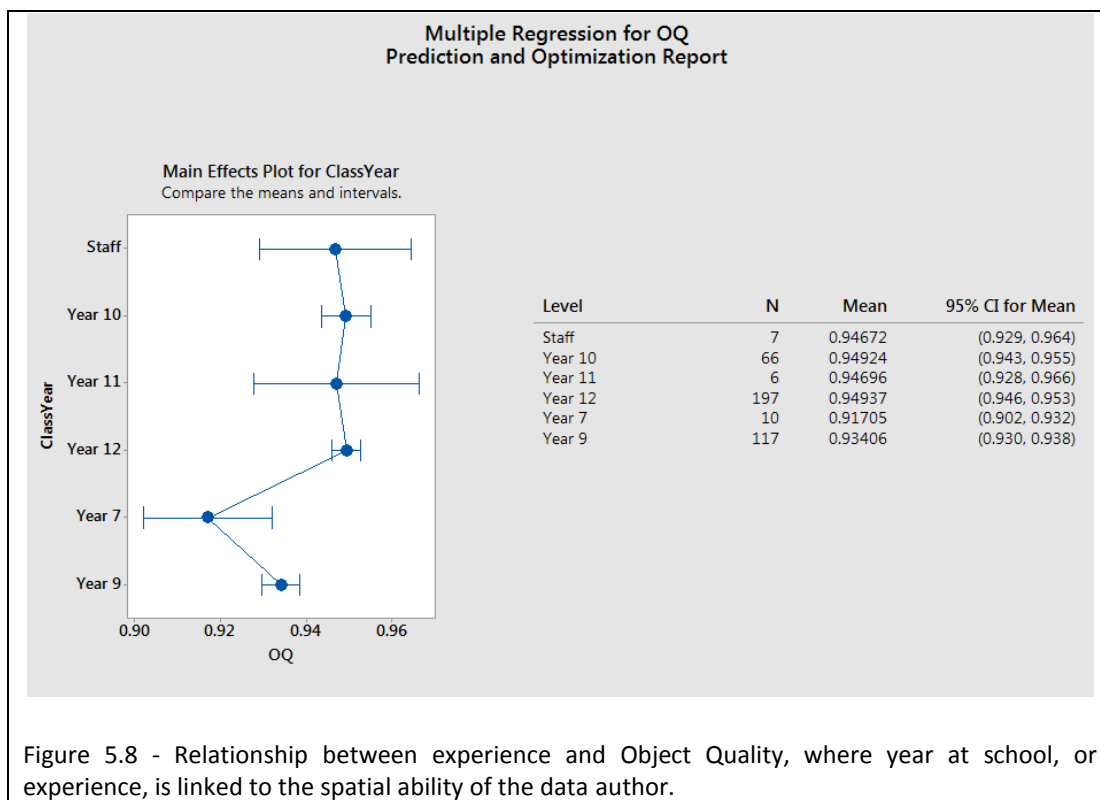
It was also interesting to note the performance of technical subjects such as Art and Graphics exceeded that of Mathematics, which was in direct contrast to recent studies. It would be useful to conduct a further analysis on this data using a larger reference dataset, in order to provide a more reflective sample size and robust analysis. For the purposes of the VGTrust model, however, these data prove an alignment between crowdsourcing reality and the relevant literature of the association of spatial ability with academic discipline. The impact of these results is discussed further in Section 6.2.



5.6.2 Experience

This study used a contributor's year at school as the way to quantify experience in terms of the VGTrust model. The assumption being tested was that an increase in a person's age would show a correlation with an increase in OQ of any given feature in the sample dataset.

The sample result in most cases matched the expected trend, with "Staff" scoring highly (OQ mean = 0.94672) and younger students showing relatively poorer spatial ability (Year 7 OQ mean = 0.91705). Year 9 students also achieved a lesser OQ mean, and Year 12 students scored a higher OQ mean (0.94937), which exceeded that of participants in the "Staff" category. An unexpected result was the high performance of students in Year 10, a relatively younger year level at secondary school. These results are depicted in Figure 5.8. The sample size for Year 10 ($n = 66$) is large enough to ensure that this result is statistically significant. The relationship between experience and other factors related to the data author, in particular subject expertise, are canvassed in detail in Section 6.2.



5.6.3 Activity Space

Activity space was measured using two metrics. First, the Euclidean distance between the centroid of the crowdsourced feature, and the coordinate representing a participant's home suburb was measured. Second, the distance between the centroid of the feature and the coordinate representing the participant's school was calculated. These two measures were used to assess the potential contribution of the "activity space" theory proposed by Goodchild (2007).

A linear regression equation was used to model the association between OQ and each activity space respectively. The regression for the school activity space showed that there was no statistically significant relationship between OQ and the distance of a feature from a participant's school, and that this measure explained less than 1% of variance in OQ. The data collected for the school component of activity space exhibited a clustering that was actually more indicative of discrete data rather than the continuous information that is actually represented. This result is further expanded on in the Discussion chapter, although it is likely that this trend is caused by the large catchment size of high school zoning, insofar as these include a large number of city suburbs, where participants were more likely to concentrate their activities.

This result can be contrasted with the relationship between OQ and distance to a participant's home suburb. The linear regression equation revealed a strong statistical association between OQ and nearness to a contributor's home suburb, or activity space. Although only a small percentage of the overall model could be explained by this measure, there was a clear relationship with a p value of $p < 0.001$. The data distribution shows a strong clustering of data with a high OQ within 6km of a person's home suburb centroid. This not only indicates a preference for contributors to map their own activity space, but also revealed a quadratic trend showing a decline in OQ with an increase in distance from one's home suburb. The results indicate an inversely proportional relationship between OQ and distance from suburb, which corresponds with the original activity space theory (Goodchild 2007). Figure 5.9 depicts the results for these two measures, and indicates that a person's home suburb is a better indicator of activity space than the area where they work, in this case a participant's school area. As noted above, the clustering revealed by the analysis of the distance to schools is perhaps suggestive of a geographic scale that is too coarse to identify trends. A school catchment could include up to 15 suburbs, and also travel to and from school would not necessarily follow a straight line. Future research should examine the effect of the transport network on activity space calculations. Crowd participants demonstrated a tendency to map areas around their houses there is a natural trend for higher quality in these areas. This phenomenon will be explored further in Section 6.2.

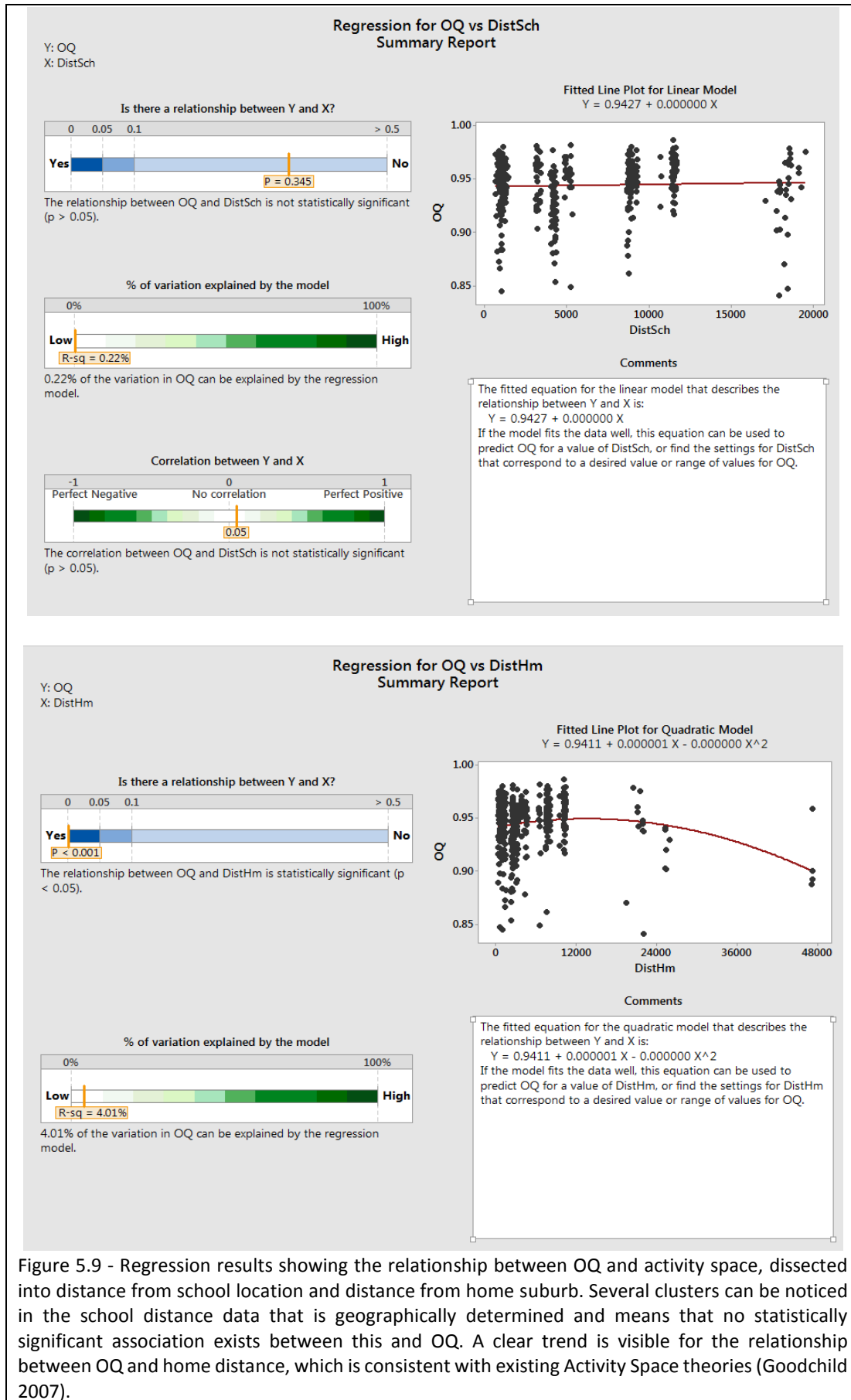


Figure 5.9 - Regression results showing the relationship between OQ and activity space, dissected into distance from school location and distance from home suburb. Several clusters can be noticed in the school distance data that is geographically determined and means that no statistically significant association exists between this and OQ. A clear trend is visible for the relationship between OQ and home distance, which is consistent with existing Activity Space theories (Goodchild 2007).

5.6.4 Spatial Precision

Three factors influencing spatial precision were assessed in the “Building Our Footprints” implementation of VGTrust. These were capture scale (or zoom level), an assessment of the number of vertices in each polygon, and through an iterative measure of the internal angles present within each polygon.

It was difficult to statistically analyse the effect of capture scale on the quality of the crowdsourced features, given that participants were instructed during data capture to improve their quality by zooming in further. For this reason, 95% of the data points sampled for further analysis were captured at the most precise scale, 1:250, resulting in insufficient data to create any inference of statistical relevance.

The second measure of spatial precision was an assessment of the number of vertices in each polygon. This measure was a case specific rule, and worked on the assumption that a larger number of vertices would be associated with a greater level of detail for a feature. This association could be particularly true in the case where overhanging eaves are present over doorways, porches and other protrusions, as opposed to a basic shape, such as a rectangle. This concept is by its very nature linked to the scale of capture, as it is impossible to capture such detail without viewing the imagery at a scale that makes these features visible. The greatest value in vertex counting was the ability to make a binary distinction between a building and a polygon that could not represent a building. In the physical world it is extremely unlikely that a building will be visible as a triangle when viewed from above. It is also recognised that a large proportion of building footprints, particularly older houses and garages, are simple shapes that generally only contain four vertices, therefore a vertex count of four was used as an initial binary cut off to infer the capture of a building from a non-building, or an accidental or erroneous capture.

Using four vertices as a minimum for building analysis, the number of vertices captured per footprint was then compared to the OQ measure generated. The analysis revealed a statistically significant relationship ($p < 0.001$) between vertex count and OQ, and was represented by a quadratic model, as depicted in Figure 5.10. Of interest is that an increase in vertices improves OQ to a point, before beginning to have a negative effect on quality. The analysis suggests that OQ improves with the inclusion of additional vertices, until peaking at 24. From that point additional vertices appear to have no tangible effect on OQ, until the number of vertices exceeds 30, when a negative effect is revealed. The binary nature of building specific analysis means that the minimum number of four vertices may have contributed to the high clustering of OQ values in the sample, as all geometrically incorrect features were filtered by this test, and contributors were encouraged to further edit these features in order to achieve a minimum standard. Nevertheless, the data show that there is a correlation between the number of vertices present and the quality of a crowdsourced feature.

The final measure of spatial precision was a measurement of the internal angles of a building polygon. Again, this measure was case specific to building polygons, and

worked on the assumption that a building is unlikely to contain internal angles of less than 90 degrees, or more than 270 degrees. This measure generated no statistical relationship to OQ, and may be more appropriate as a binary filter running in a separate layer of processing to the primary VGTrust model. As no statistical relationship was identified, this measure is not analysed further in this section.

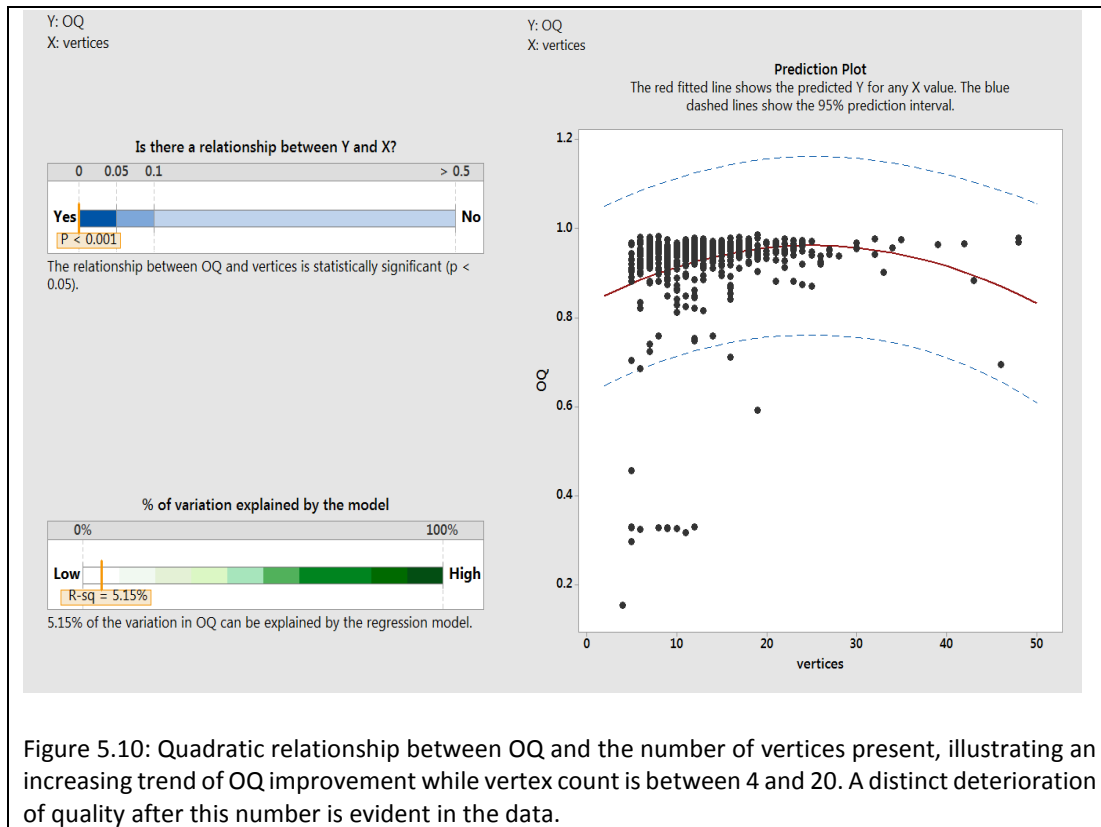


Figure 5.10: Quadratic relationship between OQ and the number of vertices present, illustrating an increasing trend of OQ improvement while vertex count is between 4 and 20. A distinct deterioration of quality after this number is evident in the data.

5.6.5 Temporal Quality

The final individual quality measure assessed in the building footprint case study was temporal quality, or change over time. This was tangibly measured by Linus' Law, the number of edits made to any given feature, and can be used to illustrate concepts of crowd control and peer endorsement of quality.

The results for this test were inconclusive, with a linear regression analysis identifying no statistically significant relationship between OQ and the number of edits ($p = 0.564$), as depicted in Figure 5.11. Indeed, while the trend line shows only a very minor improvement to maximum OQ score as numbers of edits improves, what the data distribution also illustrates is a marked decrease in lower OQ values with an increase in feature edits. This pattern suggest that although the maximum OQ score is unaffected by the number of edits to that feature, the number of features achieving higher OQ values increases proportionally with the number of edits, and by the fourth edit, there are no features achieving comparatively low OQ

values. This pattern will be illustrated further during response optimisation for the factors in the model.

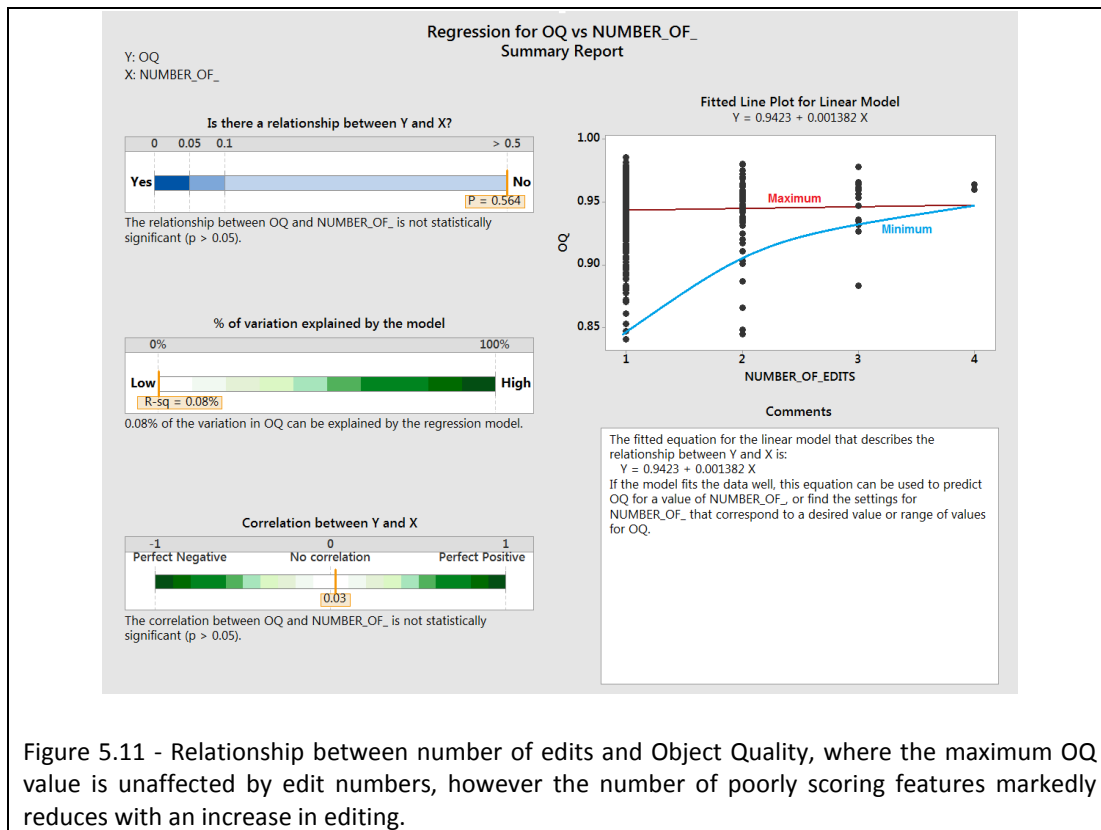


Figure 5.11 - Relationship between number of edits and Object Quality, where the maximum OQ value is unaffected by edit numbers, however the number of poorly scoring features markedly reduces with an increase in editing.

5.6.6 Relationship of Modelled Components to Each Other

A Principal Component Analysis (PCA) was carried out on the seven variables measured and tested within the model, in order to understand if any relationships existed between these variables. As the PCA deals with numerical data, the school subject and school year variables were coded according to the effect that they had on the OQ result, as defined in Sections 5.6.1 and 5.6.2. These values were coded in ascending order of impact, so that the PCA could interpret an increase in number as an increase in effect on OQ. For example, in terms of expertise, or preferred school subject, Geography was coded as a “1”, as its impact on OQ was the least significant, while Biology was coded as a “7”. The same process was followed for experience, or year at school, in terms of the results obtained by the study. The results of the PCA are shown in Figure 5.12.

A PCA is designed to detect significant associations between variables, and the interpretation of its results first require a somewhat arbitrary decision regarding the significant of the correlation. In this case, an association has been deemed significant if the factor loading of coefficients exceeds 0.5 in either a positive or negative direction. This threshold is generally considered an appropriate level by statisticians (Urdan 2010). The closer a factor loading value is to zero, the less correlated the variables are assumed to be. Therefore, any factor loading value

smaller than -0.5 or greater than 0.5, was deemed to show a significant association between variables.

Eigenanalysis of the Correlation Matrix

Eigenvalue	1.8532	1.6179	1.0914	1.0024	0.7603	0.4853	0.1896
Proportion	0.265	0.231	0.156	0.143	0.109	0.069	0.027
Cumulative	0.265	0.496	0.652	0.795	0.904	0.973	1.000
Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7
CAPTURE_SC	0.044	-0.086	0.462	-0.851	0.024	0.223	0.045
NUMBER_OF_	-0.034	0.234	-0.678	-0.444	-0.516	-0.119	-0.087
YearCode	-0.323	-0.630	-0.130	-0.099	-0.009	-0.401	0.558
SubjectCode	-0.454	-0.494	-0.196	0.035	-0.010	0.535	-0.474
DistHm	0.531	-0.280	-0.012	0.184	-0.480	0.492	0.366
DistSch	0.512	-0.462	0.056	-0.048	-0.075	-0.458	-0.551
vertices	0.379	-0.081	-0.518	-0.176	0.705	0.192	0.126

Figure 5.12 - Principal Components Analysis for factor association in the modelled variables.

Principal Component 1 (PC1) showed an equal positive association between both activity space variables – the distance a feature is from its creator’s home and from its creator’s school. The values of 0.531 and 0.512 respectively indicate that either one could positively affect the other if changed. This is an interesting result given that the distance from a participant’s school was shown to have no statistically significant impact on OQ through the regression analysis, and can be explained by the geographic phenomenon of school zoning, where a person’s home suburb is inexorably linked to the geographic location of their school.

Principal Component 2 (PC2) shows a negative relationship between year at school and preferred subject, in terms of the impact that these factors had on OQ. Year at school is weighted more heavily than subject (-0.630 and -0.494 respectively) which indicates that year at school is a strong indicator of preferred subject, and therefore spatial ability. This suggests that as participants gained experience through the school system, or by getting older, and were able to specialise in terms of their subjects, there more senior students who were involved in subjects associated with a higher spatial ability. These students were more likely to participate and achieve high OQ values in mapping activities. This result supports the concept of spatial ability being associated with certain subjects, and with the notion that experience in a particular discipline will also enhance a contributor’s spatial ability.

Principal Component 3 (PC3) showed a negative relationship between the number of edits to a feature and the number of vertices. When there are fewer edits to a feature, there are also fewer vertices captured. The factor loading on the “number of edits” variable is heavier than on the “number of vertices” variable, which suggests that it is the number of edits that has a direct impact on the number of vertices contained in a feature. This is consistent with crowdsourcing theory and Linus’ Law, which suggests that a feature will display greater precision with the more edits it receives, and the number of vertices contained within a feature has

been linked to an increase in OQ. Interestingly, as noted above, both Linus' Law (number of edits) and number of vertices, follow a specific quadratic curve that suggests the greatest impact is made on feature quality with the first 13 – 20 instances, and negligible improvement is evident thereafter. This result is also consistent with the theory underpinning Linus' Law and Geometric Quality (Haklay et al. 2010).

Principal Component 4 (PC4) shows a directly causal relationship between the capture scale of a feature and the number of edits that feature receives. With a heavy factor loading of -0.851, capture scale is deemed to be the factor that determines the extent of this relationship, i.e. capture scale is the cause and the number of edits is the effect. Number of edits is shown to decrease as capture scale decreases (noting that a decrease in capture scale is actually an improvement in the precision – 1:250 is a higher resolution scale than 1:2000). This result shows a clear and expected effect of scale in a digitising application, and suggests that if a feature is captured at the best possible scale, then fewer edits are required to lift that feature to a high standard of quality.

Principal Component 5 (PC5) shows that the number of vertices in a feature has an effect on the number of edits made to that feature. With a factor loading of vertices = 0.705, and with number of edits = -0.516, this shows that if a feature is captured with a higher number of vertices in the first instance, there is a lesser need for subsequent edits in order to achieve a good standard of OQ.

Principal Component 6 (PC6) shows a positive correlation between subject code and the distance between a feature and its author's home suburb. Although the correlation between these two variables is weaker than in other principal components, this may suggest a tendency for participants with a greater spatial ability to correctly map the area in which they live, which further reinforces the accepted theory of Activity Space (Goodchild 2007).

Principal Component 7 (PC7) reveals an equal but opposite association between year at school and features captured near to that school. This shows that as the school year of a participant increases, the distance of their feature capture from their school decreases. This is suggestive that younger students prefer to map their home area, while older students, who have had a longer association with their school and therefore a better knowledge of the area, are more inclined to engage in the mapping of that area. This PC is linked heavily to PC 1 and PC 2, and suggests that all elements of the participant are linked in some way, including experience, expertise, and activity space.

5.8 Optimising the VGTrust Model

A response optimised regression model was run to predict the level of each variable that contributed most significantly to a high OQ score. All seven previously assessed variables were modelled, and the results are described below in terms of the three major components of VGTrust.

5.8.1 Attributes of the Data Author

The four attributes of data author optimised were experience, expertise, distance from home, and distance from school. The latter two factors representing the activity space of the participant. The response optimised regression showed that all four variables were associated to each other and had a statistically significant relationship to OQ. The four variables combined accounted for 19.40% of the variance in OQ, which is 10% less with that previously ascribed in the theoretical VGTrust model. The difference is explained by the nature of OQ, as OQ is a measure strictly of geometric quality, and not attribute quality. The optimised regression results, displayed below in Figure 5.13, suggest that trust in the data author account for nearly 20% of the geometric quality of a feature. These attributes were original weighted at 30% in the VGTrust model, and from these results the difference of 10% accounting for attribute correctness may be appropriate, but was not tested in this case study.

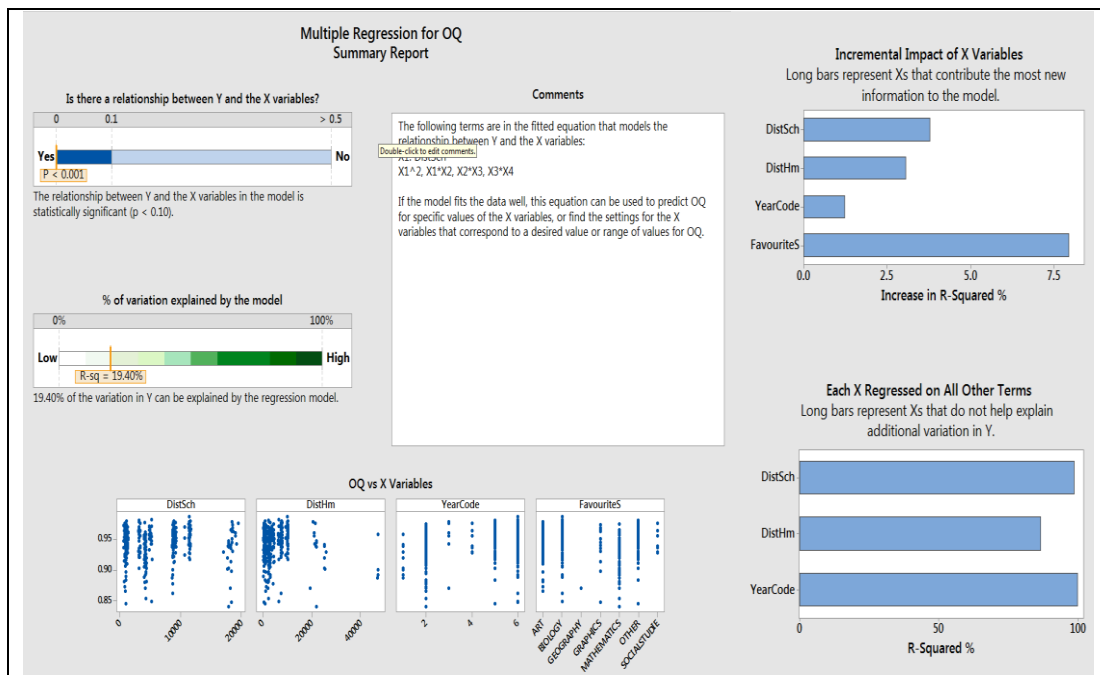


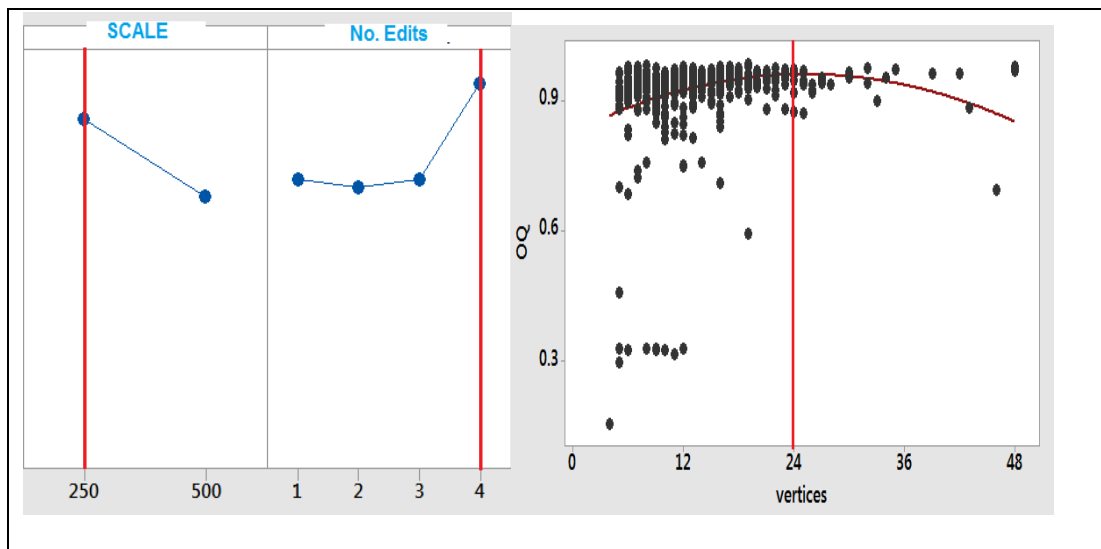
Figure 5.13 - Response optimised regression for data author model components. The intention of the regression is to maximise OQ. This assessment indicated that data author attributes accounted for 19.40% of the variance in OQ, and of the four components, expertise was more influential than activity space and experience.

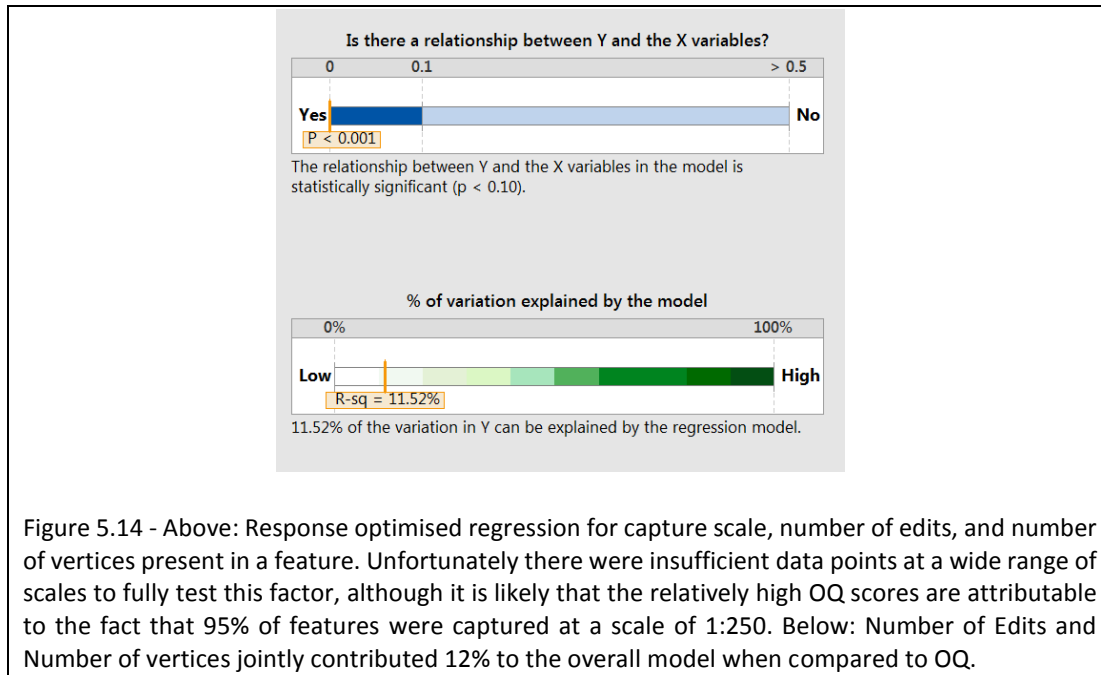
Within the “Data Author” component of the model, this analysis indicated that expertise (preferred subject) had a significantly greater effect than the other components, followed by activity space, and with experience, or year at school, making the least impact on OQ.

5.8.2 Aspects of Spatial Trust & Temporal Trust

A response optimised regression analysis was run with the remaining three variables: capture scale, number of vertices, and number of edits. There was not sufficient data to test the effect of scale on OQ, due to the way in which the data was collected and filtered. The collection process resulted in a 95% homogenous dataset, with capture being undertaken at a scale of 1:250. Nevertheless, given the high clustering of the OQ values generated, it can be inferred that the one consistent variable present for all data points was their digitisation at the maximum possible capture scale. This approach infers that, although unable to statistically test using this reference dataset, capture scale contributed a large degree to the quality of building footprint features collected. Further studies are required to explore this aspect of the hypothesis, and these would benefit from a larger reference dataset where features at multiple capture scales could be tested.

The number of vertices present and the number of edits received by a feature were found to be statistically significant in terms of their effect on OQ, and were collectively assessed to contribute 12% toward the overall score of a feature’s OQ. The optimised response suggested that maximum OQ could be gained at a capture scale of 1:250, with approximately twenty vertices in a feature, and after each feature had been edited and refined four times. The response optimisation charts are depicted in Figure 5.14.





5.8.3 The Whole VGTrust Model

The response optimised regression results indicated that the three component groups of the VGTrust model all contributed to the identification of quality for crowdsourced geographic data. These results indicated that author attributes can account for 20% of the spatial precision of a feature, while temporal and some specific spatial elements accounted for 12%. Attribute quality was unable to be assessed with the sample data in this study. The remaining 60% - 70% of the trust model is likely to be inferred by capture scale and other spatial attributes, particularly because the data from this case study was digitised rather than captured in the field.

With the analysed data now informing the further development of the VGTrust model, the following chapter will explore how the originally proposed model can be refined and appropriately deployed for crowdsourced data that is both digitised and primarily collected in the field. Chapter 6 will assess these factors against their underpinning theory, explore these results further and propose reasoning for their distribution. This will ensure a robust model is presented that will promote the assessment and acceptance of crowdsourced geographic data into authoritative datasets.

6: Discussion

Seven key measures of trust were examined in Chapter 5, these being the data author's expertise, experience, and activity space, as well as each feature's unique spatial descriptors – capture scale and number of vertices present. Additionally, a measure of provenance – Linus' Law – was examined in this case study. Each of these measures will now be discussed in terms of their relationship to existing academic thought. Gaps in the model will also be identified, and a number of future research options will be proposed that will enhance and augment the base VGTrust model derived for crowdsourced geographic information.

The trustworthiness of crowdsourced geographic information has been a subject of close scrutiny since Goodchild coined the term "Volunteered Geographic Information" for these data in 2007. The existing literature in this area examines crowdsourced geographic information through various lenses, by looking at elements of the data's spatial precision and geometric correctness (Bishr and Mantelas 2008), aspects of its history or provenance (Mooney et al. 2010), as well as how aspects of the author of that data could be used to infer conclusions about its quality (Goodchild 2008, Elwood 2008). Each of these individual categories of trust predictors have been examined in other studies, although to date none exist where all three general elements of trust have been combined to generate a single metric that could inform reliable re-use of the data. Such a metric is recognised as being of value, particularly to authoritative mapping organisations as a way to augment spatial data supply chains.

This research proposes a single VGTrust model which combines three such measures of trust as attributes of the data author, its spatial precision, and aspects of its temporal quality. The VGTrust model's purpose is to allow authoritative mapping organisations to easily assess crowdsourced geographic information with a view to integrating it into authoritative map products and supply chains. The "Building Our Footprints" mapping competition was the first implementation of this model, and the results from the data generated, as presented in Chapter 5, illustrated both similarities and differences between the existing literature and this case specific implementation.

6.1 Integrating Crowdsourced with Authoritative Data

In a traditional top-down mapping paradigm, agencies responsible for the production and dissemination of authoritative geographic data have placed a significant emphasis on the precision, quality, and ultimately the reliability of the data they produce. This top-down approach is historically the domain of government agencies or national mapping companies, and relies on a labour intensive process of extracting, verifying, and validating data from a range of

sources, most of them authoritative in their own right. An example of this protocol or approach in New Zealand is the production of the Topographic 1:50,000 map series (Topo50) by Land Information New Zealand, a process that utilises data sources, such as NASA supplied satellite imagery or terrain information. This process is also time consuming, and leads to a risk that the authoritative product produced at its conclusion is in fact out of date. A similar example is one of New Zealand's fundamental datasets (<http://www.linz.govt.nz/about-linz/our-location-strategy/geospatial-strategy-and-work-programme/fundamental-geospatial-data>), the NZ Land Cover Database (LCDB). The magnitude of the dataset, and the relative scarcity of authoritative source data, means that any given feature within the dataset could be up to ten years out of date. In other words, the reliance on authoritative data sources and strict validation processes leads to a highly precise authoritative product, however this assumption of correctness is only valid for the point in time at which the data was created. This gives rise to an obvious question – if a dataset is out of date, can it truly be considered reliable and therefore authoritative?

Mapping agencies are asking similar questions. In order to keep pace with rapidly evolving and mobile technologies, as well as the consumer expectation for real time data that comes with these developments, agencies are re-examining their data supply chains (Clouston 2014). In order to ensure that their products have more currency and relevance, mapping organisations are looking to CGI or VGI, but continue to grapple with the issue of maintaining the reliability, or even *perceived* reliability, of their products – in essence their authoritativeness.

There are eight characteristics of a dataset, authoritative or otherwise, that are widely perceived to indicate its inherent quality. These are Lineage, Positional accuracy, Attribute Accuracy, Logical Consistency, Completeness, Semantic Accuracy, Usage, Purpose and Constraints, as well as Temporal Quality (Van Oort 2009, Haklay 2010). When considered together, these measures can be seen as ways to assess the quality of a complete dataset. The VGTrust model is proposed as a way to assess the reliability of crowdsourced geographic data, in order to allow its seamless integration into an existing authoritative dataset.

With the integration of authoritative and crowdsourced data in mind, there are a number of factors that lose relevance to the model outcome. For example, a measure of dataset completeness is not required when the purpose of the activity is to add or update individual features to a collection that was complete to begin with. In essence, the benefit of crowdsourcing is to provide currency where certain aspects of a dataset may be out of date. Equally, when choosing to crowdsource a particular feature type (or types) for a given dataset, the purpose and constraints related to that schema and its collection are particularly relevant, as well as its lineage and temporal quality. Semantic accuracy is a key component of any assessment of crowdsourced data, although this has not been considered further in this research. A brief discussion of semantic and ontological considerations will follow in Section 6.7, although the domain constraints associated with facilitated crowdsourcing means that an effective trust model can be established without the need to consider these components further.

This research focussed on a subset of quality measures for crowdsourced geographic information, in order to allow a means to infer trust in these data and allow their integration in the supply chain for authoritative data and mapping agencies. The use of the VGTrust model will allow an agency's product to increase in currency without losing its authoritativeness, or most importantly, its *reputation* as authoritative.

6.2 Trust in the Data Author – Model Performance

Significant research has been conducted to examine how trust in an author of data can be used as a proxy for trust in crowdsourced data (See Chapter 2), through social network analysis, reputation and connection mapping, and the examination of existing facilitated crowdsourcing solutions, such as OSM (Coleman et al. 2009, Keßler & de Groot 2013, Golbeck et al. 2008). Golbeck et al. (2008) discussed reputation, or the “Small World Theory” as a means to infer quality. This theory at its heart is based upon social network analysis, and relies on peer endorsement of skill as a way to define a contributor's level of expertise. It works primarily in a facilitated environment such as OSM, where there is a means to provide a feedback loop, as multiple participants within the application can endorse another's contribution, or vice versa (Gira et al. 2009). This feedback loop acts to build or deconstruct a participant's reputation as able, and is also present and functioning in cases such as New Zealand's “Trademe”, an online trading application (www.trademe.co.nz).

Gira et al. (2009) divide reputational trust into two facets - credibility and expertise. These have been measured in this study through the assessment of each participant's year at school and favourite subject, which were provided for in the model in terms of the work of Shea et al. (2001). Their study posited a connection between a person's spatial ability and their preferred academic subject – in the case of this thesis, at secondary school – although the concept was tested and proven sound at all levels of education (Wai et al. 2009).

Both Shea et al. and Wai et al. identified a strong association between spatial ability and subjects such as mathematics, technical drawing, and the physical sciences, with a lesser but still positively correlated relationship with geography and the social sciences. No relationship was identified with subjects such as physical education or the performing arts.

The data collected through “Building Our Footprints” contained a mixture of different disciplines, and as such a number of conclusions were able to be drawn that reinforced the Wai et al. and Shea et al. theories. Seven separate subject categories were provided for in the data, although 66 features assessed were associated with the subject “OTHER”, and therefore did not contain sufficient detail for further investigation.

The remaining five subject types – Art, Biology, Geography, Graphics, Mathematics, and Social Studies – resulted in quality associations that were for most subjects consistent with the conclusions drawn by previous studies. The physical sciences were under-represented in the sample, with only Biology featuring from that field. Biology was the most represented subject in the sample ($n = 180$), and therefore its results when compared to OQ have the smallest confidence interval. As outlined in Chapter 5, Biology was the highest performing subject overall, followed in descending order by Art and Graphics, although this latter subject attracted a larger confidence interval on the basis of its smaller sample size ($n = 17$). These results are consistent with the findings of Shea et al. and Wai et al., as all three of these subjects are representative of those in “tier one” as described in their research, where subjects were grouped into “tiers” according to their influence on the spatial ability of students. Tier 1 subjects were most closely associated with spatial ability, Tier 2 to a lesser degree, and Tier 3 showed little or no corresponding spatial ability.

Social Studies and Geography were described in the literature as “second tier” subjects – displaying some correlation with spatial ability, but to a lesser degree than subjects contained in “tier one”. The results from “Building Our Footprints” affirmed these conclusions, with these subjects showing a lesser association with OQ, although with a particularly small sample size ($n=1$), the result for Geography is unable to be assessed further in this study.

The most surprising result revealed in this research was the relative underperformance of Mathematics when compared to its subject counterparts, scoring below all other subject associations bar Geography. This finding is surprising, given that both the Shea and Wai studies most strongly associate mathematics with spatial ability. This finding certainly warrants further investigation, and would benefit from an assessment of the entire population of building footprint data ($n = 18,792$). This assessment was not made as a reference dataset for comparison did not exist to intersect with all 18,792 features.

A possible explanation for the relative underperformance of Mathematics is the nature of the competition process itself, which was strongly cartographic. The digitised environment, combined with dynamic and high resolution colour imagery, as well as a classified symbology regime meant that the environment was more relevant to the artistic side of geography – cartography – and less akin to the more abstract spatial concepts found in the study of mathematics at secondary school.

Based on relationships with favourite subjects, the VGTrust model as originally proposed has been significantly endorsed, subject to a refinement that would see Mathematics placed in the second tier of subject associations. The principal component analysis (PCA) identified that this measure of expertise was strongly associated with both experience and activity space, the other variables assessed under the category of “data author” within the model, an association that will be discussed in depth following an explanation of the other two variables. The exploratory regression analysis showed that these three variables, in combination, account for approximately 20% of the overall variation in the model.

The second element measured to model trust in the data author was experience, a valuable component of Golbeck's definition of reputational trust (Golbeck 2008). This study tells that the more time a data volunteer has spent in a given field, using a particular skillset, the more able that person will be in that field. In the case of this research, expertise was measured through school subject association, therefore *experience* in terms of the model was deemed to be the length of time each participant has been at that school. Making the assumption that most participants started at the same school in Year 9 as they finished in Year 13, the measure of this experience was classified accordingly.

Student participants were drawn from Years 7, 9, 10, 11 and 12, with some features also contributed by school staff members. The results generally supported the theory that an increase in age would represent an increase in OQ, although the data for Year 11 showed a surprising low point when compared to Year 10. OQ results for Staff were also surprising lower than those for Years 10 and 12, although both of these anomalies could be attributable to low sample sizes ($n = 6$ and $n = 7$ respectively) and subsequently high confidence intervals.

Interestingly, the PCA showed a direct relationship between both the *experience* and *expertise* variables within the model, which further supports the theories proposed by Golbeck et al. (2008), Golbeck & Hendler (2004) and Van Exel et al. (2008). These theories state that *experience* and *expertise* are complimentary and combine to present an overall reputational trust measure. Until this point, an assessment of reputation trust for the creation of VGI has been largely theoretical, with some studies quantifying reputational trust by social network theory through functional examples – notably the “TrustBot” email assessor described by Golbeck & Hendler (2004).

The implementation of the “VGTrust” model through “Building Our Footprints” has further quantified these propositions, and corroborated their findings. VGTrust has demonstrated a direct relationship between the quality of a crowdsourced geographic feature and the expertise and relevant experience of its creator. Reputational trust, therefore, is a valuable component when assessing trust in crowdsourced geographic data.

It seems logical that this concept could be further extended to incorporate more complex measures of reputational trust, and each implementation of the VGTrust model would necessarily require a case-specific calibration in order to determine which person attributes are available for assessment within the model. VGTrust was designed for use in facilitated crowdsourcing applications by large data-handling agencies, and in such an environment the requisite information could be collected at any point when the contributor engages with the process. This was the situation with “Building Our Footprints”, with the necessary information was collected about each participant upon registration for the competition. In data harvesting applications such as this, it is of course always necessary to balance privacy concerns with the need for accurate data assessment and model integrity. The model can always be calibrated according to the limitations of the available data,

and can therefore be extended into more organic crowdsourcing applications, such as OSM.

A valuable potential source of reputation information could be through LinkedIn, a professional networking application that is used by a huge number of professionals worldwide, over a variety of different professions. The key function of LinkedIn that makes it useful for extending this research, is its ability to collect peer endorsements about an individual's expertise, based on their reputation. This is a real world implementation of Golbeck's "Small World Theory", and would be ideally suited for inclusion in the VGTrust model. Unfortunately, at the time of implementation, the API library for LinkedIn was not sufficiently mature to allow the extraction of this required information, and this source of data was therefore not available for study. Such an application would be a valuable source of data for future research.

The final element of author trust tested by "Building Our Footprints" was the concepts coined by Goodchild as a person's "activity space" (Goodchild 2008). This theory predicts that a crowdsourced feature captured near an area where its author spends significant time, or is intimately familiar, will be inherently more accurate with greater spatial and attribute correctness than one captured (mostly digitised) in an unfamiliar area. In the unique case of spatial information, the length of time a person has spent in a given geographical area will be directly and positively correlated with that quality of their data contribution about that area.

The "Building Our Footprints" application tested this concept by collecting information about the school that each participant associated with, and their home suburb. These locations were assigned a coordinate based on their address point and polygon centroid respectively, and the Euclidean distance between the crowdsourced feature and these activity space locations was compared to OQ.

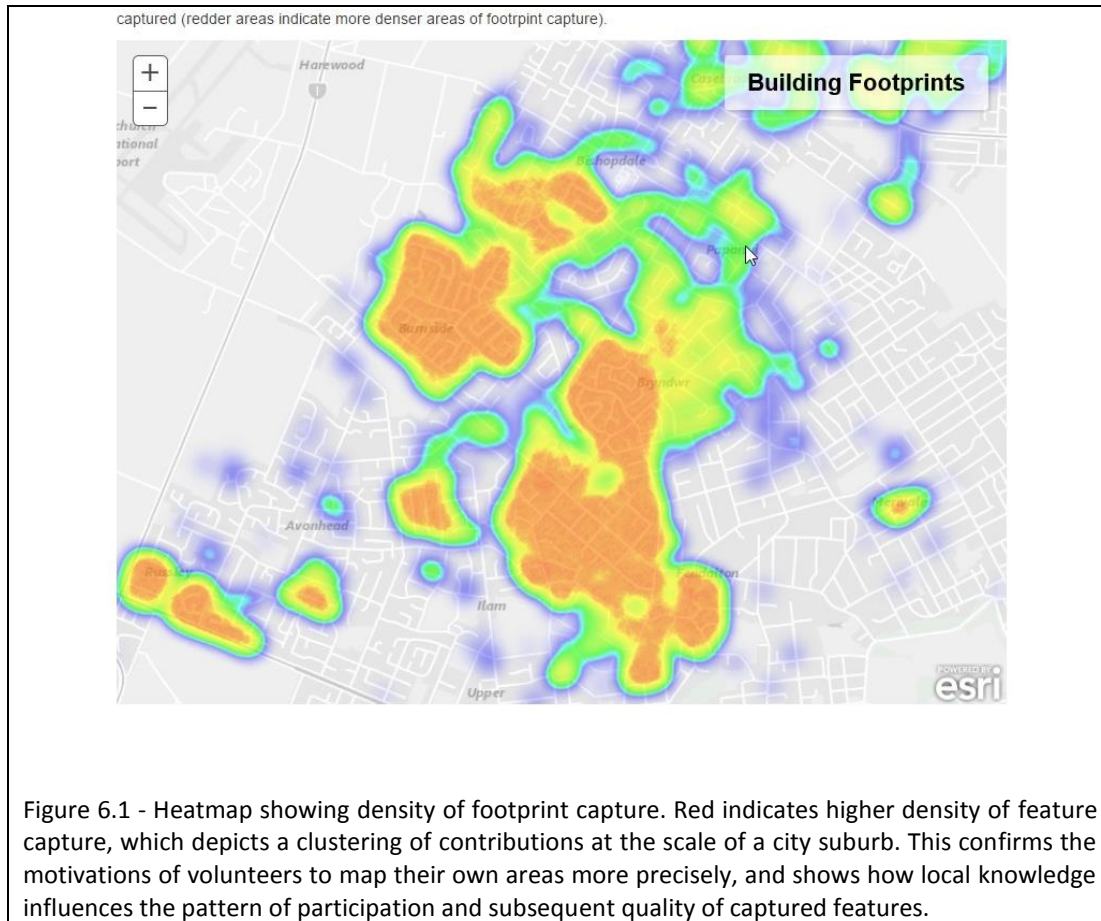
The Activity Space results when applied to participants' home suburb endorsed the theory that the quality of crowdsourced geographic data deteriorates with increased distance between that data and its creator's activity space. However, the "Building Our Footprints" case studied also revealed inconclusive results for the test when applied to school associations as a measure of activity space. These results revealed no statistically significant correlation between OQ and the distance between a participant's school location and the feature location. An inversely proportional relationship was identified when this test was applied to a data author's home suburb.

Two possible reasons may lead to this discrepancy. First, in a socio-spatial context, the relationship of the data to arbitrary administrative boundaries must be considered. For the vast majority of secondary schools in New Zealand and in many places around the world, enrolment is based upon a catchment area, or 'school zone', which is in turn linked to the suburb in which a person dwells. School zones incorporate a number of suburbs, which by their nature are smaller units of geographic division. The lack of correlation between OQ and a school-based activity space measure could simply therefore be attributable to the fact that a school

catchment is a much coarser unit of measurement, and that defining the activity space by suburb is a more appropriate scale for this assessment. Similarly, a second possible explanation is tied to engagement and motivation theories for crowdsourced applications. Coleman et al. (2009) and Heipke et al. (2010) attempt to classify participation in crowdsourcing applications into classes of motivation. One such motivation is the desire to see one's own area of interest – or activity space – mapped precisely and accurately, based on their own (and perceived superior) local knowledge.

The “Building Our Footprints” data revealed that the majority of building footprints were captured in tight clusters (Figure 6.1). These clusters, as well as an assessment of the organic growth of the dataset, showed that participants naturally began mapping in their own suburb. Although not expressly collected, anecdotal evidence suggests that each student participant began by mapping the buildings at their own address, then naturally expanded their participation into the immediate surrounding area. This area would grow organically until subsequent milestones were reached, such as the complete mapping of a street, followed by a block, followed by a collection of city blocks. It follows that this clustering is therefore more closely associated with suburb distribution than school distribution, and this is reflected in the results generated.

The PCA also shows a direct correlation between both activity space measurements, which is not surprising given the relationship between suburb and school zoning administration. Interestingly, the PCA also depicts a relationship between the home activity space measure, and the number of vertices in, and number of edits made, to a feature. The vertex and edit number measures (Linus' Law) will be discussed in detail in sections 6.3 and 6.4, as specific spatial and temporal trust variables. At this point it is important to note the correlation between these components and the Activity Space measure, as they are interconnected and supportive of each other. Described simply - the closer a crowdsourced feature is to its author's activity space, the fewer edits it will require to meet a certain accuracy standard, and the greater number of vertices, or modelled precision, it will possess. This vertex pattern confirms the notion that when a geographic volunteer is contributing data in their local area, that person's local knowledge has a positive impact on the quality of their contributions. These results indicate that the “Activity Space” theory is sound when assessing trust in CGI, and reinforces the conclusions drawn in previous studies (Goodchild 2008, De Longueville et al. 2009).



The three components of author trust – expertise, experience, and activity space – have been modelled as contributing to approximately 20% of the total variation in the model, 10% less than the original VGTrust weighting of 30%. This is likely due to the fact that OQ accounted for spatial aspects of data quality, while author trust can also be linked to attribute correctness, a factor not considered further by this study. The respective weightings of Expertise, Experience, and Activity Space were also refined from those previously proposed, with Expertise now occupying a significant proportion of the Data Author component. It can therefore be confirmed that in a facilitated digitised crowdsourcing application, the following weightings should apply:

- Expertise = 10%
- Experience = 4%
- Activity Space = 6%

These weightings indicate that 20% of the total VGTrust model should be built upon data author attributes. It must be noted that these proportions have been rounded to the nearest whole number, in order to supply future users of the model with a metric that is easy to interpret. The remaining 80% of the VGTrust model can be explained by the following spatial and temporal measures of trust.

6.3 Performance of Specific Spatial Indicators

Three aspects of spatial precision were proposed by this research, and tested through the “Building Our Footprints” application. The assessment of spatial variables were important in the case of geographic information, as they represent the uniquely spatial aspects of these data, which are not present in other crowdsourcing applications. Given the purpose for which geographic information is created and consumed, the quality of its position in space and its geometric correctness are vital aspects of its overall quality, or trust. Data consumers rely on its positional correctness for self-location, amenity-location, and navigation, among other purposes.

The three spatial variables tested in the VGTrust case study were the capture scale of a feature, the measurement and categorisation of a building polygon’s internal angles, and the number of vertices present in a building footprint feature. The first measure is generic to all digitising applications for crowdsourced data, while the latter two measures are case-specific measures related to the building footprint data type.

The scale at which each feature was captured and subsequently edited was arguably the most important measure. The original model hypothesis, upon which the case study was run, set this parameter as comprising 60% of the overall trust assessment. This is a large proportion of the VGTrust model, although the results illustrated that the accuracy and spatial precision of the polygons decreased exponentially with a reduction in capture scale.

The second element was building footprint specific and involved the measurement of each polygon’s internal angles. This angle measurement recognised that the majority of building footprints are comprised of a number of typical internal angles, particularly at their corners. The most common are 90°, 270°, and 130°. The way in which this aspect of the model functioned is outlined in detail in Section 3.3.2. No results for this aspect of the model were available from the case study, however as a case specific trust measure it is essential to note its contribution to “Building Our Footprints”.

Finally, the number of vertices were counted in each feature to predict its level of precision. The underpinning concept is that a building must have a minimum of four vertices, as it is very unlikely that a building footprints would be in the shape of a triangle. Furthermore, it was mooted that a greater number of vertices would be associated with a higher level of detail afforded to each footprint, and should be closely tied to both a building’s capture scale and distribution of internal angles. These measures were identified through a PCA as being closely related in their influence of OQ. The number of vertices associated with a feature, as well as its capture scale, were taken forward for further analysis, where capture scale was assessed as having a significant impact on the number of vertices present in a feature. This result is consistent with the proposition of De Longueville et al. (2009), that in a digitised context, scale does matter.

As noted Section 5.7.4, it was difficult to statistically analyse the effect of capture scale on the quality of the crowdsourced features. Observationally, however, the features captured at a more coarse scale were those that were significantly larger, and were not representative of buildings, such as parks and school fields. The theory of quality linkages to capture scale remains sound, as it is not possible for a human participant to view and therefore capture a building with any detail or precision at more coarse scales. Irrespective of screen size, which is in some way relevant in a digitising scenario, a building is not distinguishable as more than a basic shape at any scale greater than 1:2000. Even at a scale of 1:1000, critical details can be omitted as they cannot be distinguished. Figure 6.2 depicts this effect, showing the capture of a building, first at a scale of 1:1000, then compared at a scale of 1:250. The resolution of the imagery is such that the polygon appears more precise at the more distant scale, but at a closer resolution presents key details that were omitted from the original capture.

Furthermore, the scale at which a feature was captured and edited was proven to have a dramatic impact on overall OQ, and indeed on other variables within the model. Capture scale in particular had a dramatic effect on the number of edits associated with a feature, and was equally closely associated with the number of both vertices and edits. Principal Components 3 and 4 illustrate these dependencies, proving link between all three factors, and a strong causal effect of capture scale on edit numbers respectively. Principal Component 3 shows a relationship between number of edits (-0.678) and number of vertices (-0.518), which, based on the factor loadings, indicates that each has a similar effect on the other, although there is perhaps a stronger causal effect of edit numbers on number of vertices. Principal Component 4 shows a significant causal effect of capture scale on the edit numbers (capture scale = -0.851, number of edits = -0.444), which shows that as capture scale reduces (when a participant is zoomed in more closely), the number of edits required to bring a feature up to the desired quality level drastically reduced. The performance of this measure – Linus’ Law – is discussed further in Section 6.3.

Despite being unable to statistically analyse the effect of capture scale, its role in the overall model for a digitised example is clearly significant. Subsequent optimisation of the model identified that a scale of 1:250, the most precise available, is most likely to generate features of a higher quality. These results are consistent with the De Longueville et al. (2009) principles.



The second spatial quality measure analysed was the relationship between the number of vertices of a building feature and OQ. The sample studied showed a statistically significant relationship between both, and indicated that the number of vertices accounted for 5.15% of variance in the overall model. For ease of interpretation this figure, when rounded down to 5%, is entirely consistent with its original proposed weighting within VGTrust.

Several means of assessing spatial correctness of a crowdsourced feature were identified in the existing literature. These included the concept of “Fuzzy Distance” (Kuhn 2007), the spatial homophily of data (Bishr & Mantelas (2008) – in essence the similarity of different contributor’s interpretation of the same feature – as well as a number of other specifically spatial measures outlined by De Longueville et al. (2009).

“Building Our Footprints” placed an emphasis on the completeness and logical consistency of the dataset, which are two factors identified by Haklay (2010) as influencing the quality of VGI. In this case, the purpose of the data collection was to leverage a particular crowd in order to generate as many individual footprints as possible, and therefore the opportunity was lost to have many participants contributing information on the same feature. This meant that the “fuzzy sets” theory (Du et al. 2012) was unable to be tested, where a number of different data submissions representing the same real world feature are aggregated to find the most common, or likely, positional accuracy. Each individual data feature could then be rated according to its similarity to the aggregated ideal, and returned in a

feedback loop to the VGTrust model. Future analysis in this area could add weight to this form of crowd endorsement, albeit a passive version.

It is important to also note that the facilitated nature of “Building Our Footprints” meant that all participants digitised features using the same tools and the same base imagery. This removed a significant amount of ambiguity usually associated with generic crowdsourcing applications, or data collected in the field using an instrument and subsequently volunteered. Given the purpose of the VGTrust model – to inform integration of facilitated VGI into authoritative datasets – the case study provided valuable insight. Should the model be extended in the future to take in other modes of data collection, the choice of base imagery and capture device would potentially need to be built into the formula. Such additions could include device metadata or metadata on base source imagery including date captured, projection, and source. Such future work would be a valuable additional module for the base model.

Equally important is the examination of what is *not* present in the results – a significant level of variation in capture scale. During the facilitated “Building Our Footprints” data collection, participants were actively encouraged and directed to digitise features at the maximum possible capture scale. This guidance was given and justified as a part of the engagement strategy for the competition, so that the crowd – in this case secondary school students, would maintain a sufficient level of interest and involvement in the activity. It is therefore likely that the unusually high OQ results revealed by the results sample, and described in detail in Chapter 5, is due to this one factor, as a reduction in capture scale had a major impact on the final VGTrust result. As participants were directed to adjust the capture scale to improve their trust scores, there was little variation in the results to support a comprehensive analysis of its effect on OQ. There was however a clear relationship between capture scale and other variables, particularly the provenance factor – number of edits, or Linus’ Law, discussed below in Section 6.4.

In total, the following rounded weightings were proposed for a generic digitising application for CGI:

- Capture Scale	=	60%
- Specific Spatial Measure 1	=	5%
- Specific Spatial Measure 2	=	5%
Total effect on VGTrust of spatial variables		= 70%

6.4 The Effect of Data Provenance

A large number of methods for assessing trust based on data provenance have been proposed and discussed in *Section 2.3 – Temporal Quality* (Haklay et al. 2010, Du et al. 2012, Mooney et al. 2010, Aragón et al. 2009). Of these, an examination of Linus' Law was assessed as a part of the "Building Our Footprints" case study by examining the non-linear relationship between the number of edits received by a crowdsourced feature, and its overall quality or trust level (Haklay et al. 2010). Through their study of OSM data, Haklay et al. determined that the majority of improvement to the data was made during the first six edits to a feature. A declining rate of improvement was seen between seven and thirteen edits, with negligible change in evidence beyond that.

A similar assessment by this study showed a similar trend, although with an overall fewer number of edits. The rules of the building footprint competition required participants to achieve in excess of a particular trust rating – 75% - in order to achieve a point towards their total score. Most features achieved this rating within four edits, and, as such, this number is the highest recorded for the analysis. The trend revealed is of interest, given that a regression analysis showed no statistical relationship between the number of edits and OQ. There is very little change to the upper OQ values regardless of how many edits that feature has received. There was a significant clustering of high OQ values in both the first and second edit iterations, which therefore illustrated very little effect of this variable on the final trust scores. The effect was not statistically significant.

Upon visual examination of the data, however, there is a marked reduction in features achieving a poor OQ by the time a third and fourth edit are made. This is in itself an endorsement of Linus' Law, albeit in a marginally different context to that identified by Haklay et al. (2010). This study shows that the number of feature edits has a significant impact on the quality of the dataset as a whole, although their effect on the quality of each individual feature was less obvious. One reason for this trend is the impact of capture scale on the temporal variable. The PCA results showed a direct causal relationship between capture scale and the number of edits made to a feature (Capture Scale -0.851, Number of Edits -0.444), indicating that the scale at which the feature was captured was the primary cause of subsequent edits made to that feature. On this basis, temporal measures of trust were proposed as composing 10% of the overall VGTrust model.

Further measures of temporal quality were noted but not included in the case study for technical reasons. One of these measures, proposed by Aragón et al. (2009), suggests that a changes to a feature over time represent improvements and therefore result in a higher available trust level for that data. This *change ratio* also proposes that no edits to a particular feature over a period of time can equally be seen as an endorsement by the crowd of its quality, and should be treated as such. An examination of Aragón et al.'s *change* and *contribution ratios* would be ideally suited to a future OSM implementation.

The results obtained through this implementation study led to a finalisation of weightings in the generic VGTrust model, and further informs which information must be stored in a crowdsourcing data structure to allow an assessment of trust to continue for future data reuse.

6.5 VGTrust – Final Variable Weightings

On the basis of the data collected through “Building Our Footprints”, final weightings were established for each of the variables within the VGTrust model. The weightings were established using a regression comparison to Object Quality, an independently assessed measure of trust in a crowdsourced feature. The results largely supported the original estimates for variable weighting established prior to deployment in case study. These weightings are generic, although specific to a digitising application, and are illustrated in Figure 6.3. The weightings are summarised using the following equation:

$$VGTrust = \sum \{ (0.04 \times Experience) + (0.10 \times Expertise) + (0.06 \times Activity Space) + (0.60 \times Capture Scale) + (0.10 \times Other Spatial) + (1.10 \times Number of Edits) \}$$

Or in summary:

$$VGTrust = \sum \{ 0.20(Author) + 0.70(Spatial) + 0.10(Temporal) \}$$

Where *Author* includes elements of expertise, experience, and an assessment of activity space; and *Spatial* includes the digitised capture scale, as well as case specific spatial business rules, an example of which is the vertex count and internal angle assessment for building polygons; and *Temporal* includes elements of data provenance.

This way of assessing trust is effective where a facilitated environment exists, and the requisite attributes can be collected and stored at the time of capture. Such a scenario assumes that the organisation collecting the data has a robust way of storing it, and of protecting the privacy of the contributors who volunteer personal details about themselves to make the model function.

In many crowdsourcing scenarios, however, this auxiliary data is not present. The following section will explore smaller derivative trust assessments, using one or more components of the overall VGTrust model. It will also explore the most effective data structure for storing the attributes required to run the model, and how the need for these must be offset by privacy considerations when collecting personal data.

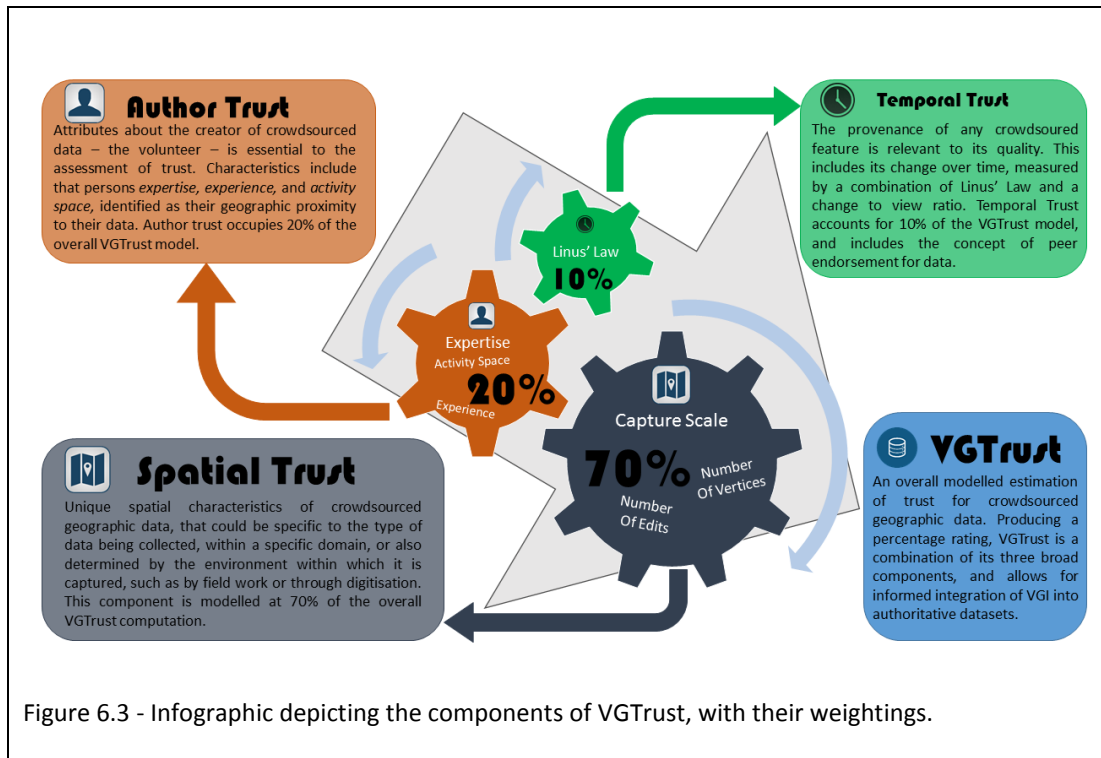


Figure 6.3 - Infographic depicting the components of VGTrust, with their weightings.

6.6 An Effective Data Structure to Model Trust

This section will focus on the optimal data structure to ensure the VGTrust model can accurately run, by providing enough metadata about a feature's capture while still maintaining its author's privacy. It has been identified that neogeographers have few skills, nor the requisite experience, to be populating what is traditionally considered to be a dataset's metadata (Johnson & Sieber 2013, Poore & Wolf 2013). It is therefore proposed that the information required to enable VGTrust is stored in the primary data structure of a feature, and will therefore travel independently with that feature, irrespective of whether or not it remains a part of its original dataset. In this context, the term "metadata" is used to illustrate the types of auxiliary information discussed by this research, such as information about a data author or that data's provenance.

Goodchild et al. (2007) posited the notion of a "Geo-Atom", which in essence is the most basic form taken by geographic data. It included a location (again at its most simple this would be a coordinate, or point type geometry), and an attribute about that data (theme), which has a particular value at a particular location. A Geo-Atom is represented by the form

$$\langle x, Z, z(x) \rangle$$

In the wider geospatial world of Web 2.0, the geo-atom potentially represents the extent of a significant proportion of data. Where data is not facilitated to collect all of the required attributes, the VGTrust model as outlined by this study would lack sufficient inputs to produce a meaningful result. Should this form of data be required for assessment, a logical example of which could be a point marker indicating the presence of a feature such as a public toilet, then other forms of quality assessment must be undertaken. The only option available in these cases is the aggregation approach, or “Fuzzy Sets” theory, where the precise location of a feature is modelled by a centroid calculation from all like features, as well as inferring attribute correctness when a high number of contributors identify the same feature within a similar location. Grira et al. (2012) and Van Exel et al. (2011) discuss the limitations of this approach, however, with the rise of what can be described as a “pack mentality”, or blind trust in other contributed data, by virtue of the fact that there is a large volume of it. A contributor will endorse the capture of a feature based on the fact that another contributor has also endorsed it, rather than on the basis of its actual correctness. If examining all of the factors that have had a demonstrated effect on the quality of a crowdsourced feature, it is logical to suggest that an assessment made without all of the components should attract a low trust rating, although not necessarily a low quality rating.

A feature, captured and assessed in this way, could easily be entirely accurate and suitable for re-use. In this situation, without the proper checks, this type of feature would simply be an unknown variable and should be treated as such. Such an approach logically involves the examination of semantic concerns, and the application of volunteer ontologies, in order to determine which VGI data in fact represent the same real world phenomena as one another. Such considerations are beyond the scope of this research, however are discussed in Section 7.1.

More mature crowdsourcing platforms, such as OSM or Wikimapia, as well as facilitated crowdsourcing applications implemented by national mapping organisations or government agencies, will have the necessary infrastructure to securely collect and maintain all of the metadata required for the model. The “Building Our Footprints” application, for example, was developed and deployed through a collaboration of central and local governmental agencies. The data structure of the application was such that it was able to store and maintain separate tables within its database for participant information, which was held securely behind established firewalls and security protocols. Such an environment also exists for applications, such as OSM, and in all cases of facilitated crowdsourcing, all of the metadata should be maintained against the feature, but could promote the final VGTrust *rating* as its key attribute. By doing so, subsequent discoverers and re-users of the data will be able to make a rapid and effective assessment as to its quality and potential fit for their purposes. This corresponds with the fitness for purpose assessments identified by Van Oort et al. (2006) and Haklay (2010), and suggests a deviation from the complex data structure presented by de Longueville et al. (2009).

Fundamentally, a data structure to represent trust in a crowdsourced feature must be sufficiently detailed but also sufficiently simple to allow its future discovery and

re-use. Any new data structure also needs to support what Poore and Wolf (2013) identify as its key requirements – to enhance its usability, ease of discovery, and the relationship between data and metadata. This thesis supports the concept identified by Sui et al. (2012), that information about the creation and subsequent development of crowdsourced data should maintain transitivity with each individual feature. Given the diversity of contributors to a crowdsourced dataset, this information cannot be aggregated to that higher level, but must remain at the level of each feature.

There are of course privacy considerations – the contributor’s precise location should not be carried with that person’s data, for reasons of personal security, but also for the sake of perception and engagement. A participant in collaborative mapping will be less inclined to continue their contributions if they feel that their personal data or privacy has been in some way compromised. Whether privacy can be actually compromised or not, the result is similar – the discouragement of active future participation in mapping. On this basis, an option to consider is the aggregation of location, or even a banding and processing of location data before association with a feature. For example, the data could indicate that the participant was within, or identified with, a certain radius of the contributed feature, and could be stratified in terms of less than 1km, between 1km and 5kms, between 5km and 10km, etc.

Nevertheless, Goodchild et al.’s “Geo-Atom” must necessarily be expanded into a “Geo-Molecule”, in order to sufficiently capture the data required for VGTrust. Such a “Geo-Molecule” could be modelled as follows:

$$\langle x, Z, A, S, T, VGT, z(x), l(A), e(A), S(f), T(f), T(ts), VGT(f) \rangle$$

Where x and Z are defined in terms of Goodchild et al.’s “Geo-Atom”, as “a point in space-time, a property (attribute), with $z(x)$ being a particular value of a property at that point.” (Goodchild et al. 2007). The additional characteristics of the data structure transform that atom into a “Geo-Molecule”, and incorporate the relevant metadata to inform trust. These would be carried against each feature not as metadata in its most pure sense, but as feature specific attribute data.

The additional parameters for a “Geo-Molecule” would be formed from the following attributes, which have been discussed in depth in the preceding sections of this chapter:

1. VGT – The finalised indicative trust rating for the feature (f), based on other parameters. This would necessarily need to be read in conjunction with an associated explanation on how it was derived, in order to properly inform its re-use. This field contains the output of the model, and is the most important aspect of its use.
2. A – Author components. This component has been further subdivided into the author’s location ($l(A)$), and the author’s expertise ($e(A)$). For non-facilitated

crowdsourced data, it is likely that these fields would attract a NULL value, as the information to populate these would not be available.

3. S – Spatial components of a feature. Such a measure is extremely case specific. It is logical that this would include the most fundamental of spatial measures, such as device precision for a field-collected feature, or capture scale for digitised data. Any domain-specific business rules, such as the internal angle measurement for a building footprint, could be implemented as a separate modular plug-in for the model. Furthermore, the aggregation of like-features captured to model the same real-world entity could be considered here.
4. T – Temporal components of a feature. On its most basic level this could include the number of edits that feature has received - Linus' Law – or data provenance. This includes a timestamp (T(ts)) for each edit, and would allow more complex analysis to be conducted subsequently, such as an assessment of the *change* and *contribution ratios* described by Aragó et al. (2009).

This geo-molecule presents a base-level means of capturing trust parameters against a data feature, which will ensure this information is both complex enough to illustrate trust, and simple enough for neo-geographers to understand and use. The VGTrust geo-molecule will further ensure that all required trust information maintains transitivity with individual features.

6.7 Limitations of VGTrust

The VGTrust model was developed as a way to allow the integration of crowdsourced geographic data into authoritative datasets. It includes a broad range of tests that model trust in the quality of that data, but is by no means exhaustive in its measures, rather focussing on what could be described as a “base platform”. This means that a number data aspects are considered when deriving an inferred quality, all of which are deemed to be fundamental to the derivation of crowdsourced quality. These measures are not without caveat, however, and could also benefit from additional modules to model further parameters.

The limitations to VGTrust can be categorised into two streams – limitations of the existing base model components, and omissions from the current model.

6.7.1 Limitations of the Base Trust Model

VGI is distinctly spatial by its nature, which carries with it a set of particular validation requirements that are not always relevant for more general crowdsourced information. Consequently the final weighted model has a notably high proposed weighting for those parameters that model distinctly spatial components of the data. Such spatial assessment measures could be inferred or actual, such as the capture scale of a building feature (inferred trust by proxy)

versus a precise measurement of its internal angles (actual), but in all situations are domain, or case, specific.

This notion leads to one of the major limitations of the VGTrust model approach – it opens the user of the model to subjectivities that may affect the final trust result, which will in turn affect the result of any analysis that has relied on that data at a particular quality level. For example, the use of case specific parameters is predicated on the notion that the model user will have an inherent understanding of what makes a “good” feature of that data type, and what the meaning of geometric correctness may be in each case. That user, or super-user as the case may be, will then need to build an appropriate means of testing for that correctness, and apply that measurement within the VGTrust model. In essence, as the VGTrust model has been designed to facilitate the integration of CGI and authoritative data, there is an implicit assumption that the user of the model will have expertise and experience handling geographic data and information. It is acknowledged that despite this intent, an actual user of VGTrust may not always be so proficient.

VGTrust is also limited by its hunger to consume personal metadata about a feature’s author and the means by which it was captured. While a data structure has been proposed that could yet appropriately manage these data, the unfortunate reality is that these characteristics are not available for a large proportion of crowdsourced data.

Spatially, the model must also be calibrated to assess the scale of the dataset, and the diaspora of participants. For example, in “Building Our Footprints”, the activity space measure was calibrated on the basis of the geographic extent of the competition being a city, and stratified according to the perceived familiarity of participants within these bounds. This was further compounded by participant demographic – the classification of Activity Space was made with volunteer age in mind – secondary school students most likely possess a greater familiarity of their own area than adults, and a reduced familiarity with the wider city. This could be for a number of reasons, including a reduced need to travel large distances on a daily basis, as well as differing means of transportation. Whilst an adult in the workforce may traverse a city one or more times during the course of a normal day, a student would most likely travel to school or after-school activities on a daily basis, all of which are likely to be in a similar geographic area to their main domicile – limiting the range of their knowledge but in turn deepening it. Adolescents would also be more likely to travel using modes of transport that encourage observation of their surrounds – walking or cycling, or on public transport. This often results in a heightened awareness of the detail of a local environment, detail that may not be noticed by somebody in the same area who is focussed on negotiating traffic on the local roads.

A number of “reduced” versions of the model were discussed in earlier sections of this chapter, however the degree upon which any result from these can be relied is questionable. This research has shown that in order to appropriately model all elements of trust for this particular type of information, all of the VGTrust input measures must be present and considered together. Such a conclusion suggests

that VGTrust is in itself only fit for a select purpose, to facilitate the integration of crowdsourced data into authoritative datasets, and may generate unreliable results when put to use by other users within a crowdsourcing ecosystem. In a facilitated collection environment, the appropriate metadata could be collected upon registration, as was seen during the “Building Our Footprints” competition, which would allow the model to run at least once with all of the required inputs. Any data generated could of course be re-used on the basis of its trust rating, and potentially added to at a later date. This perhaps suggests that the VGTrust model should not be run if a certain threshold of data inputs is not met.

6.7.2 Omissions from the Base Model

VGTrust presents a general way in which to quantify trust in crowdsourced data, and allow the appropriate deployment of quality assurance resources when integrating these data into authoritative datasets. Its design is modular, and recognises that this research presents only a broad picture of the fundamentals of trust for this type of data. There are of course many ways that the model could be enhanced, although almost all rely on the notion that additional data and metadata will be available to feed these components. These range from more complex ways of measuring change over time, through to the inclusion of semantic measures and a study of volunteer ontologies.

The examination of semantic issues is a large topic that could form the basis of an entire study in itself. As such, an investigation into its use for trust modelling has been limited to an exploration of its potential, and should be treated as an opportunity for future research in this area, in particular as a modular addition to the base VGTrust model.

Volunteer ontologies are extremely diverse, and in the case of online mapping tools, likely to be geographically dispersed. By examining the intent and description of a captured feature, an understanding of semantics could not only enhance trust in a features attributes, but also allow a more accurate assessment of its spatial precision. This could be described as spatial homophily, the “Many Eyes” principle, or an aggregation of location. The concept is simple in its intent – with a greater number of features capturing the same real world phenomenon, a precise spatial location and definition of that feature can be gleaned by aggregating all of the volunteered interpretations of that particular ‘thing’. The difficulty is defining exactly which ‘thing’ is which, and can be aided by an investigation of semantics.

By knowing how a particular feature is described by a certain user group, and how these definitions relate to those of other user groups, a picture of likeness can be established, with those features subsequently used for other spatial analyses. A basic example is the case of rugged or hilly terrain. A person who normally dwells in a mountainous region may describe a particular terrain as “hill country”, whereas a person who lives in a flat coastal region may describe the same terrain as “mountainous”. The difference is a matter of perception rather than fact – the hills

remain the same size despite naming convention – and it is these differences in perception that could be modelled through semantic trust.

The same approach could logically be extended to attribute correctness. Is a river a river, or is it a stream? Is a particular water body a pond, or a lake? The differentiation becomes particularly important in large cities, where there could be any number of different data features within a relatively tight geographic area. Line features on a map could depict railway lines, monorail or tram lines, and all would be presented spatially almost on top of one another. How could these be differentiated? What is light rail – a tram track or a train track? A study of semantics and volunteer ontologies could go a significant way to clarifying some of these questions, and allowing more precise spatial aggregation.

6.8 Summary

The results and analysis of data collected during this study, particularly through the “Building Our Footprints” application, showed a direct correlation with the existing literature regarding the quality of CGI and VGI. Through a statistical examination of the three broad categories of trust – Author, Spatial, and Temporal – the VGTrust model has been calibrated to provide a generic algorithm that can be used in a range of CGI and VGI applications (depicted in Figure 6.4). Most importantly, this study has proven that an automated metric can be established to allow the integration of CGI into authoritative datasets, particularly through facilitated applications and environments.

The VGTrust model is founded on the principles of Web 2.0, and has been proven by case study to be appropriate to integrate CGI into authoritative datasets. As discussed in Chapter 7, a future study and implementation of semantic and ontological trust would, however, transform a Web 2.0 informational assessment into the realms of Web 3.0 as volunteer intelligence.

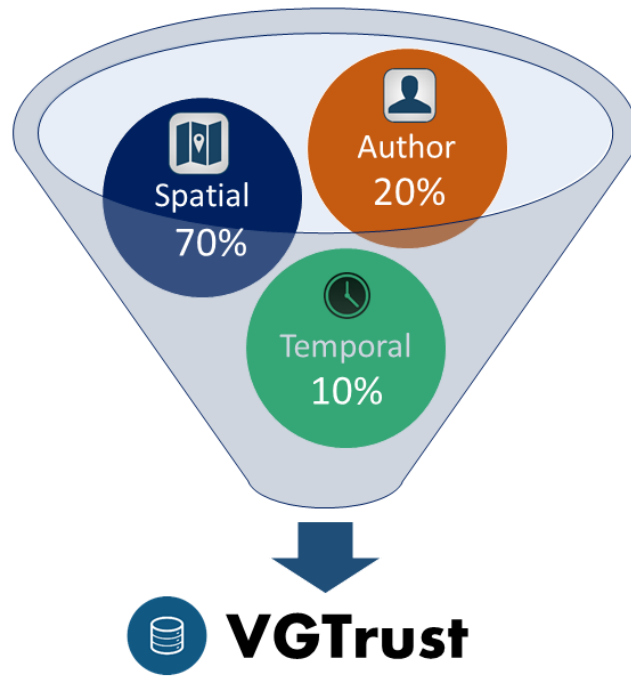


Figure 6.4 - Overview diagram of VGTrust.

7: Conclusion

Crowdsourced, or volunteered, geographic data represents a valuable source of multi-sensory, near real-time information about the world in which we live, and the way that people and societies interact. Its most powerful attribute is currency, and its attraction is economy – every citizen in western society records aspects of their daily activities in some way with a personal sensor – most commonly a mobile phone, tied to a location. This data is essentially “mass-produced” by a “crowd” of mappers as a by-product of simply being – ostensibly for no associated cost.

Crowdsourcing as a source of data is attractive to authoritative mapping agencies, with a traditional “top-down” mapping paradigm involving rigorous quality assurance processes, the need for highly trained professional staff, and high associated overhead costs. Traditional data or map products are therefore relatively highly priced, and based on data accurate only at a given point in time. Authoritative mapping agencies have to date been unable to trust the quality of CGI to the extent that it could be included in their data supply chains. Although much work on CGI has been dedicated to assessing certain aspects of quality, until now there has not been a way to holistically measure these data for an overall picture of trust.

This research presented VGTrust, a model that combines aspects of author trust, spatial trust, and temporal trust into a simple metric that generated an overall inference of quality. The use of VGTrust will give authoritative mapping agencies the opportunity to augment their spatial data supply chains with CGI thus significantly improving the currency of those products. Furthermore, this study deployed VGTrust in a facilitated crowdsourcing application – “Building Our Footprints” – with the support of government agencies Land Information New Zealand and Environment Canterbury Regional Council. The results of “Building Our Footprints”, and the subsequent analysis of data, demonstrated that VGTrust is a viable option for future crowdsourcing applications by government. The model successfully estimated trust, and therefore quality of, the CGI tested.

Additionally, this thesis proposed a data structure that could appropriately store the requisite information for assessing trust. This data structure, a geo-molecule, ensure that all the information required is captured against, and travels with, each individual feature. It is not metadata in the traditional sense of the term, rather an extension of basic CGI schema, and its implementation is vital for the ongoing transitivity of trust through CGI user communities.

This study found that the three broad categories of trust – author, spatial, and temporal, were essential when assessing trust in CGI. This largely coincides with expectations based on previous work, although some of the individual factors varied in their importance in the final model compared to the weight with which they were originally proposed.

Trust in a data author was measured in three ways, through an assessment of author expertise, experience, and activity space. Expertise and experience were the two component of what most scholars term “reputation”, while the theory of activity space was deemed essential to any assessment of data authorship. Activity space theory is inherently linked to the motivations for participation of volunteer mappers, a phenomenon which was observed through the cluster analysis of “Building Our Footprints” data. The data revealed that mappers did prefer to focus their contributions around their own activity spaces, a pattern that became evident as the crowdsourced dataset grew organically outward from these centres over time. Despite this observation, activity space accounted for a smaller weighting in the final model than originally expected, as did all aspects of data authorship.

Of the two reputational characteristics tested, expertise was weighted higher than experience when proposed originally in a theoretical model. These two characteristics combined were weighted equally with Activity Space, at 15% each, together accounting for 30% of VGTrust. These weightings proposed activity space as the most significant contributor to author trust. The results of this research, however, led to an alteration of all of these components, identifying the expertise of a data author as the primary measure of trust, followed by activity space, and to a lesser extent experience. Additionally, the results suggested that author trust only accounts for 20% of the overall model, a reduction from the earlier hypothesis.

Conversely, spatial trust elements, originally weighted at 60% of VGTrust, gained a further 10% significance in its final weighting. The “Building Our Footprints” application involved digitising of geographic information, which always suggested that spatial trust would play a principal role in overall quality assessment, so the high weighting was not surprising. The results of this study did however illustrate the importance of case-specific parameters for trust assessment, specific to feature type or domain. It was found that these specific business rules could act as a Boolean test to determine whether a feature is within its required domain. This is particularly important in facilitated crowdsourcing applications.

Finally, temporal trust was assessed as making up 10% of VGTrust. This was consistent with the weighting originally proposed, and reflected an endorsement of the Linus’ Law concept (Haklay et al. 2010). This concept states that a non-linear but incremental relationship exists between the number of edits to a feature and the quality of, or trust in, that feature. There are further measures of temporal trust that were not tested here, but should form the basis of further research.

7.1 Future Research Directions

The base VGTrust model could be extended beyond the limitations of the case study and theoretical discussion in this thesis. The temporal trust component of the model could benefit from more complex measures of trust than Linus’ Law alone, and should in future iterations include the concepts of a *change ratio* and

contribution ratio as described by Aragó et al. (2009). It would be of significant value to be able to test crowd endorsement through observing a lack of change to a feature, in addition to modelling improvement through positive and visible change only.

Furthermore, the study of semantic considerations and volunteer ontologies presents a particularly important avenue for future research, insofar as in many ways it may negate the need for a facilitated approach to data collection. Should semantic attributes be fully understood and modelled, and sufficiently automated for inclusion in a trust metric, then the locational precision of a crowdsourced feature could be accurately determined by aggregation, or the Many Eyes principle (Haklay et al. 2010), without the need for additional metadata such as capture scale or information on device precision. A semantic assessment could also reveal differences in author age, education and location, all aspects considered by VGTrust model as author trust. If such attributes could be inferred semantically, then this removes not only the need for a complex “Geo-Molecule” data structure, but also negates the need to consider privacy concerns as a barrier to crowd participation. The VGTrust base model is sufficiently modular to allow such extensions to be incrementally added over time. In addition to this, a study of semantic considerations could reveal a way to infer trust in attribute correctness, an aspect of CGI that is of equal importance as spatial accuracy and precision. Attribute correctness has not been assessed in this research, although is identified as the next priority area for research in this domain.

7.2 Concluding Statements

This thesis proposed a means to assess CGI for trustworthiness, and inferred quality. The proposed VGTrust model seeks to assess trust holistically – to develop a single metric that is both comprehensive and easy to understand, therefore allowing authoritative mapping agencies to augment their spatial data supply chains with CGI. This research tested the question,

“Can an algorithmic model be used to establish trust for CGI, thereby facilitating its assimilation into authoritative spatial datasets?”

VGTrust is an effective way to answer this question. Building on previous research, VGTrust enables agencies to crowdsource their own data in real time, and maintain a level of confidence in the calibre of that data. Previous studies (see Chapter 2) have focused on comparing post-collection CGI with existing authoritative datasets, examining only one aspect of trust at any given time. This thesis changed that dynamic, combining different aspects of trust into one metric, and through a facilitated case study, demonstrated that VGTrust functions in a practical, commercial environment, as well as at a theoretical level. By proposing the geo-molecule data structure for CGI, this thesis has further established a way by which

trust information can endure as it flows from producer to consumer, through producers, communities, and spatial data supply chains.

This research focused on facilitated solutions to CGI capture, but has captured principles that can be applied to data harvesting, or more organic crowdsourcing applications. VGTrust will form the foundation of future research in this area, and is sufficiently modular to allow the incorporation of other extensions as the field comes of age. VGTrust is the vehicle by which crowdsourcing theory can be transported into the commercial world, realising significant cost and knowledge benefits for local, national, and global communities.

Bibliography

Aragó, P., Díaz, L., & Huerta, J. (Unpublished) A Quality approach to Volunteer Geographic Information, *Sourced from www.researchgate.net*

Bartoschek, T., & Keßler, C. (2013) VGI in Education: From K-12 to Graduate Studies, In D. Sui, S. Elwood & M. godchild (Eds.), *Crowdsourcing Geographic Knowledge, Volunteered Geographic Information (VGI) in Theory and Practice* (pp. 341 – 360), New York, Springer.

Bishr, M., & Janowicz, K. (2010) Can we trust information?-the case of volunteered geographic information. In *Towards Digital Earth Search Discover and Share Geospatial Data Workshop at Future Internet Symposium, volume* (Vol. 640).

Bishr, M., & Mantelas, L. (2008) A trust and reputation model for filtering and classifying knowledge about urban growth. *GeoJournal*, 72(3-4), 229-237.

Budhathoki, N., Bruce, B. & Nedovic-Budic, Z. (2008) Reconceptualising the Role of the User of Spatial Data Infrastructure, *GeoJournal*, Vol. 72, pp. 149-160.

Clouston, A. (2014) Crowdsourcing the Cadastre: The Applicability of Crowdsourced Geospatial Information to the New Zealand Cadastre. *Unpublished Master's Thesis*, Victoria University of Wellington.

Coleman, D., Georgiadou, Y. & Labonte, J. (2009) Volunteered Geographic Information: The Nature and Motivation of Producers, *International Journal of Spatial Data Infrastructures Research*, Vol. 4, No. 1, pp. 332-358.

Craglia, M. (2007) Volunteered Geographic Information and Spatial Data Infrastructures: When do Parallel Lines Converge? *Position Paper for the VGI Specialist Meeting*, Santa Barbara 13 – 14 December.

De Longueville, B., Ostlander, N. & Keskitalo, C. (2010) Addressing Vagueness in Volunteered Geographic Information (VGI) – A Case Study, *International Journal of Spatial Data Infrastructures Research*, Vol. 5.

Du, H., Alechina, N., Hart, G., Jackson, M., Anand, S., Morley, J., Leibovici, D., Ware, M., (2012) Geospatial Information Integration for Authoritative and Crowd Sourced Road Vector Data, *Transactions in GIS*, Vol. 16 No. 4, pp. 455-476.

Elwood, S. (2008) Volunteered Geographic Information: Future Research Directions Motivated by Critical, Participatory, and Feminist GIS, *GeoJournal*, Vol. 72, pp. 173-183.

Fan, H., Zipf, A., Fu, Q., & Neis, P. (2014) Quality assessment for building footprints data on OpenStreetMap. *International Journal of Geographical Information Science*, 28(4), 700-719.

Flanagin, A. & Metzger, M. (2008) The Credibility of Volunteered Geographic Information, *GeoJournal*, Vol. 72, No. 3-4, pp. 137-148.

Golbeck, J. (2008) Weaving a web of trust. *Science*, 321(5896), 1640-1641.

Golbeck, J. (2009) Trust and nuanced profile similarity in online social networks. *ACM Transactions on the Web (TWEB)*, 3(4), 12.

Golbeck, J., Hendler, J. (2004) Inferring Reputation on the Semantic Web, *Proceedings of the 13th International World Wide Web Conference*, May, Vol. 316.

Golbeck, J., Parsia, B., & Hendler, J. (2003) Trust Networks on the Semantic Web, Springer Berlin Heidelberg pp. 238-249

Goodchild, M. (2007) Citizens as Voluntary Sensors: Spatial Data Infrastructure in the World of Web 2.0, *International Journal of Spatial Data Infrastructures Research*, Vol. 2, pp. 24-32.

Goodchild, M. (2009) NeoGeography and the nature of geographic expertise, *Journal of Location Based Services*, Vol. 3, No. 2 pp. 82-96.

Goodchild, M. F., Yuan, M., & Cova, T.J. (2007). Towards a general theory of geographic representation in GIS. *International journal of geographical information science*, 21(3), 239-260.

Goodchild, M. (2008) Spatial Accuracy 2.0, *Proceedings of the Eighth International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*, Vol. 1, pp. 1-7.

Gira, J., Bédard, Y., & Roche, S. (2010) Spatial data uncertainty in the VGI world: Going from consumer to producer. *Geomatica*, 64(1), 61-72.

Haklay, M. (2010) How Good is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets, *Environment and Planning B: Planning and Design*, volume 37, pp 682 – 703.

Haklay, M., Basiouka, S., Antoniou, V., Ather, A. (2010) How Many Volunteers Does it Take to Map an Area Well? The Validity of Linus' Law to Volunteered Geographic Information, *The Cartographic Journal*, Vol. 47 No. 4 pp. 315-322.

Heipke, C. (2010) Crowdsourcing Geospatial Data, *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 65, pp. 550-557.

Johnson, P. A., & Sieber, R. E. (2013). Situating the Adoption of VGI by Government, in Sui, D., Elwood, S., & Goodchild, M. (2013) crowdsourcing geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice, *Springer*, New York.

Keßler, C., & de Groot, R. T. A. (2013) Trust as a proxy measure for the quality of Volunteered Geographic Information in the case of OpenStreetMap. In *Geographic Information Science at the Heart of Europe* (pp. 21-37). Springer International Publishing.

Keßler, C., Janowicz, K., & Bishr, M. (2009) An agenda for the next generation gazetteer: Geographic information contribution and retrieval. In *Proceedings of the 17th ACM SIGSPATIAL international conference on advances in Geographic Information Systems* (pp. 91-100). ACM.

Keßler, C., Trame, J., & Kauppinen, T. (2011) Tracking editing processes in volunteered geographic information: The case of OpenStreetMap. In *Identifying objects, processes and events in spatio-temporally distributed data (IOPE), workshop at conference on spatial information theory* (Vol. 12).

Kuhn, W. (2007). Volunteered geographic information and GIScience. *NCGIA, UC Santa Barbara*, 13-14

Lankes, R. D. (2008). Credibility on the internet: shifting from authority to reliability. *Journal of Documentation*, 64(5), 667-686.

Mooney, P., Corcoran, P., & Winstanley, A. C. (2010) Towards quality metrics for OpenStreetMap. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems* (pp. 514-517). ACM.

Osterman, F. & Spinsanti, L., (2011) A Conceptual Workflow for Automatically Assessing the Quality of Volunteered Geographic Information for Crisis Management, *Proceedings of AGILE*, April.

Parker, C., May, A. & Mitchell, V. (2012) Understanding Design With VGI Using an Information Relevance Framework, *Transactions in GIS*, Vol. 16, No. 4, pp. 545-560.

Pickles, P. (2011). Social dreaming in the 21st century, the world we are losing, *Psychodynamic Practice: Individuals, Groups and Organisations*, Vol. 17, No. 1, pp101-104.

Poore, B. & Wolf, E. (2013) Metadata Squared: Enhancing Its usability for Volunteered Geographic Information and the GeoWeb, in Sui, D., Elwood, S. & Goodchild, M. (2013) crowdsourcing geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice, *Springer*, New York.

Poser, K. & Dransch, D., (2010) Volunteered Geographic Information for Disaster Management with Application to Rapid Flood Damage Estimation, *Geomatica*, Vol. 64, No. 1, pp. 89-98.

Roick, O., Loos, L., & Zipf, A. (2012) A Technical Framework for Visualizing Spatio-Temporal Quality Metrics of Volunteered Geographic Information. *Proceedings of the GEOINFORMATIK 2012—Mobility and Environment*.

Seeger, C. (2008) The Role of Facilitated Volunteered Geographic Information in the Landscape Planning and Site Design Process, *GeoJournal*, Vol. 72, No. 3-4, pp. 199-213.

Shea, D. L., Lubinski, D., & Benbow, C. P. (2001). Importance of assessing spatial ability in intellectually talented young adolescents: A 20-year longitudinal study. *Journal of Educational Psychology*, 93(3), 604.

Sui, D., Elwood, S., & Goodchild, M. (2013) crowdsourcing geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice, *Springer*, New York.

Trame, J., & Keßler, C. (2011). Exploring the lineage of volunteered geographic information with heat maps. *GeoViz, Hamburg, Germany*.

Urdan, T. C. (2010) Statistics in Plain English, *Taylor & Francis Group*, New York.

Van Exel, M., Dias, E., & Fruijtier, S. (2010) The Impact of Crowdsourcing on Spatial Data Quality Indicators, *Proceedings of the 6th GIScience International Conference on Geographic Information Science*, September p. 213.

van Oort, P. A. J., Hazeu, G. W., Kramer, H., Bregt, A. K., & Rip, F. I. (2010) Social networks in spatial data infrastructures. *GeoJournal*, 75(1), 105-118.

Wai, J., Lubinski, & Benbow, C. P. (2009). Spatial ability for STEM domains: aligning over 50 years of cumulative psychological knowledge solidifies its importance, *Journal of Educational Psychology*, 101, (pp. 817 – 835)

Ye, M., Janowicz, K., Mulligann, C., & Lee, W. (2011) What You Are is When You Are: The Temporal Dimension of Feature Types in Location-Based Social Networks, *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, November, pp. 102-111.

Zhan, Q., Molenaar, M., Tempfli, K., & Shi, W. (2005). Quality assessment for geo-spatial objects derived from remotely sensed data. *International Journal of Remote Sensing*, 26(14), 2953-2974.

Zook, M., Graham, M., Shelton, T., & Gorman, S. (2010) Volunteered Geographic Information and Crowdsourcing Disaster Relief: A Case Study of the Haitian Earthquake, *World Medical Health Policy* 2, No. 2, pp. 7-33.

Websites

www.canterburymaps.govt.nz/buildingourfootprints

www.linz.govt.nz

<http://www.linz.govt.nz/about-linz/our-location-strategy/geospatial-strategy-and-work-programme/fundamental-geospatial-data>

<http://www.linz.govt.nz/about-linz/our-location-strategy/geospatial-strategy-and-work-programme/new-zealand-geospatial>

<http://www.linz.govt.nz/about-linz/our-location-strategy/geospatial-strategy-and-work-programme/what-sdi>

www.mgis.ac.nz

www.minitab.com

www.openstreetmap.org

www.tearaway.co.nz

www.wikimapia.com

Appendix 1: VGTrust.py

```
def VGTrust(featureX, featureY, homeX, homeY, schoolX, schoolY, subject, year,
zoom, vertex, pointlist, editlist=[], timeList=[]):
```

```
    import math
```

```
    homerate = 0
```

```
    distance = math.sqrt( (featureX - homeX)**2 + (featureY - homeY)**2 )
```

```
    if distance <= 1000:
```

```
        homerate = 10
```

```
    elif distance > 1000 and distance <= 2000:
```

```
        homerate = 9
```

```
    elif distance > 2000 and distance <= 5000:
```

```
        homerate = 8
```

```
    elif distance > 5000 and distance <= 7500:
```

```
        homerate = 7
```

```
    elif distance > 7500 and distance <= 10000:
```

```
        homerate = 6
```

```
    elif distance > 10000 and distance <= 12500:
```

```
        homerate = 5
```

```
    elif distance > 12500 and distance <= 15000:
```

```
        homerate = 4
```

```
    elif distance > 15000 and distance <= 20000:
```

```
        homerate = 3
```

```
    elif distance > 20000 and distance <= 25000:
```

```
        homerate = 2
```

```
    else:
```

```
        homerate = 1
```

```
    schoolrate = 0
```

```
    distance2 = math.sqrt( (featureX - schoolX)**2 + (featureY - schoolY)**2 )
```

```
    if distance2 <= 1000:
```

```
        schoolrate = 10
```

```
    elif distance2 > 1000 and distance <= 2000:
```

```
        schoolrate = 9
```

```
    elif distance2 > 2000 and distance <= 5000:
```

```
        schoolrate = 8
```

```
    elif distance2 > 5000 and distance <= 7500:
```

```

    schoolrate = 7
elif distance2 > 7500 and distance <= 10000:
    schoolrate = 6
elif distance2 > 10000 and distance <= 12500:
    schoolrate = 5
elif distance2 > 12500 and distance <= 15000:
    schoolrate = 4
elif distance2 > 15000 and distance <= 20000:
    schoolrate = 3
elif distance2 > 20000 and distance <= 25000:
    schoolrate = 2
else:
    schoolrate = 1

```

```

expertise = 0

```

```

subject = subject.upper()

```

```

aList = ['MATHEMATICS', 'SCIENCE', 'GRAPHICS', 'PHYSICS', 'CALCULUS',
'STATISTICS']
bList = ['SOCIALSTUDIES', 'ART', 'CHEMISTRY', 'BIOLOGY', 'GEOGRAPHY']
cList = ['ENGLISH', 'LANGUAGES', 'MATERIALS TECHNOLOGY', 'HISTORY',
'ECONOMICS', 'ACCOUNTING', 'BUSINESS STUDIES', 'CLASSICS']
dList = ['PYSICAL EDUCATION', 'MUSIC', 'ESOL']

```

```

if subject in aList:
    expertise = 10
elif subject in bList:
    expertise = 7.5
elif subject in cList:
    expertise = 5
else:
    expertise = 2.5

```

```

age = 0

```

```

year = year.upper()

```

```

if year == "YEAR 13":
    age = 10
elif year == "STAFF":
    age = 10
elif year == "YEAR 12":
    age = 8
elif year == "YEAR 11":

```




```

    age = 6
elif year == "YEAR 10":
    age = 4
elif year == "YEAR 9":
    age = 3
else:
    age = 2

```

```

scale = 0

```

```

if vertex <= 3:
    scale = 0
elif anglerate <= 3:
    scale = anglerate
else:
    if zoom == 250:
        scale = 10
    elif zoom == 500:
        scale = 9
    elif zoom == 1000:
        scale = 8
    elif zoom == 2000:
        scale = 7
    elif zoom == 4000:
        scale = 6
    elif zoom == 8000:
        scale = 5
    elif zoom == 16000:
        scale = 4
    elif zoom == 32000:
        scale = 3
    else:
        scale = 2

```

```

noEdits = 0

```

```

noEdits = len(editlist)

```



```
linusLaw = 0
```

```
if noEdits >= 13:
    linusLaw = 10
elif noEdits == 12:
    linusLaw = 9
elif noEdits == 11:
    linusLaw = 9.5
elif noEdits == 10:
    linusLaw = 8
elif noEdits == 9:
    linusLaw = 7.6
elif noEdits == 8:
    linusLaw = 7.4
elif noEdits == 7:
    linusLaw = 7.2
elif noEdits == 6:
    linusLaw = 7
elif noEdits == 5:
    linusLaw = 6
elif noEdits == 4:
    linusLaw = 5
elif noEdits == 3:
    linusLaw = 4
elif noEdits == 2:
    linusLaw = 3
elif noEdits == 1:
    linusLaw = 2
else:
    linusLaw = 1
```

```
vertices = 0
```

```
if vertex <= 3:
    vertices = 1
elif vertex == 4:
    vertices = 5
elif vertex > 4 and vertex <= 6:
    vertices = 6
elif vertex > 6 and vertex <= 8:
    vertices = 7
elif vertex > 8 and vertex <= 10:
    vertices = 8
elif vertex > 10 and vertex <= 12:
    vertices = 9
else:
    vertices = 10
```

```

anglerate = 0

from math import degrees, atan

coordlist = []

# Convert polygon geometry array into an iterable list

for p in pointlist:
    coordlist.append(p)

bearinglist = []

# iterate through list of point tuples to calculate the bearing between the two
points.
# push each bearing calculated into a new list.

for i in xrange(len(coordlist) - 1):
    pointA, pointB = coordlist[i], coordlist[i + 1]

    x1 = pointA[0]
    y1 = pointA[1]
    x2 = pointB[0]
    y2 = pointB[1]

    deltaX = (x2 - x1)
    deltaY = (y2 - y1)

    angle = 0

    if (deltaX == 0):
        if (deltaY > 0):
            angle = 90.0
        else:
            angle = 270.0
    elif (deltaY == 0):
        if (deltaX > 0):
            angle = 0.0
        else:
            angle = 180.0
    else:
        angle = math.degrees(math.atan((deltaY+0.0)/deltaX))
        if (deltaX < 0):

```

```

    angle += 180
elif (deltaY < 0):
    angle += 360

bearinglist.append(angle)

finalBearing = bearinglist[0]
bearinglist.append(finalBearing)

# Iterate through the bearing list and calculate the difference in angle between
# each bearing and the next bearing in the list.
# for each bearing difference calculated (angle), classify according to pre-defined
logic
# applicable to building footprints. This is modelled on the assumption that the
majority of
# angles on a building will be 90 degrees, or 180 degrees, or 135 degrees. Take
each
# base 10 rating and append to a new list.

ratinglist = []

for i in xrange(len(bearinglist) - 1):

    b1, b2 = bearinglist[i], bearinglist[i + 1]

    bdiff = (b2-b1)

    if bdiff < 0:
        bdiff += 360
    else:
        bdiff = bdiff

    result = 0

    if bdiff > 85 and bdiff <= 95:
        result = 10
        #print 10
    elif bdiff > 175 and bdiff <= 185:
        #print 9
        result = 9
    elif bdiff > 115 and bdiff <= 125:
        result = 8
        #print 8

```

```

elif bdiff > 125 and bdiff <= 135:
    result = 6
    #print 6
elif bdiff > 265 and bdiff <= 275:
    result = 7
    #print 7
elif bdiff > 275 and bdiff <= 359:
    result = 5
    #print 5
elif bdiff > 185 and bdiff <= 195:
    result = 4
    #print 4
elif bdiff > 45 and bdiff <= 60:
    result = 3
    #print 3
elif bdiff <= 45:
    result = 1
    #print 1
else:
    result = 2
    #print 2

ratinglist.append(result)

```

```

# calculate average rating for all angles in the polygon, and return this as the
result of the function
# this will provide an overall quality estimate for the feature.

```

```
totalrate = sum(ratinglist)
```

```
numrates = len(ratinglist)
```

```
anglerate = ((totalrate+0.0)/numrates)
```

```

#ratio = 0

#import datetime
#import math

#orderedtime = []
#for t in sorted(timeList):
    #orderedtime.append(t)

#start = orderedtime[0]
#end = orderedtime[-1]

#struct_date1=datetime.datetime.strptime(start, "%Y-%m-%d %H:%M:%S")
#struct_date2=datetime.datetime.strptime(end, "%Y-%m-%d %H:%M:%S")
#diff = struct_date2 - struct_date1

#seconds = diff.total_seconds()

#minutes = seconds/60

#time = math.log10(minutes)
#edits = math.log10(noEdits)

#ratio = (edits/time)*10

print homerate
print schoolrate
print expertise
print age
print scale
print vertices
print anglerate
print linusLaw

VGTrust =
((((homerate*0.25)+(schoolrate*0.25)+(expertise*0.35)+(age*0.15))*0.30)+(((scal
e*0.60)+(vertices*0.20)+(anglerate*0.20))*0.60)+(linusLaw*0.10))/10)

print VGTrust
return VGTrust

```

Appendix 2: VGTrust.sql

```

USE [GIS]
GO
/***** Object: UserDefinedFunction [dbo].[fnVGTrust]  Script Date: 2/07/2014
10:09:10 a.m. *****/
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO

CREATE FUNCTION [dbo].[fnVGTrust]
(
    @featureX float,
    @featureY float,
    @homeX float,
    @homeY float,
    @schoolX float,
    @schoolY float,
    @subject varchar(50),
    @year varchar(7),
    @zoomscale int,
    --@vertexcount int,
    @noedits int,
    @firstedit datetime,
    @lastedit datetime,
    @shape geometry
)
RETURNS float
AS
BEGIN
    DECLARE @Result float

    DECLARE @homerate float, @schoolrate float, @expertise float, @age
float, @scale float, @corners float, @linusLaw float, @ratio float, @anglemeasure
float, @timedifference bigint;

    -- Calculate the home rating factor
    SET @homerate = dbo.fnVGTrust_ActivitySpaceNE(@featureX, @featureY,
@homeX, @homeY);

    -- Calculate the school rating factor
    SET @schoolrate = dbo.fnVGTrust_ActivitySpaceNE(@featureX,
@featureY, @schoolX, @schoolY);

    -- Calculate the expertise factor

```



```

SET @expertise = dbo.fnVGTrust_Expertise(@subject);

-- Calculate the age factor
SET @age = dbo.fnVGTrust_SchoolYear(@year);

-- Calculate the scale factor
SET @scale = dbo.fnVGTrust_ZoomLevel(@zoomscale);

-- Calculate the linus law factor
SET @linusLaw = dbo.fnVGTrust_LinusLaw(@noedits);

-- Calculate the edits ratio
SET @timedifference = DATEDIFF(minute,@firstedit,@lastedit);

-- Rates the number of vertices
--SET @corners = dbo.fnVGTrust_Vertices(@vertexcount);
SET @corners = dbo.fnVGTrust_Vertices(@shape.STNumPoints());

-- Rates the angles
SET @anglemeasure = [dbo].[fnVGTrust_VertexRating](@shape);

IF @noedits <= 1
BEGIN
    SET @ratio = 0; --dbo.fnVGTrust_Vertices(@vertexcount);
END
ELSE
BEGIN
    SET @ratio = LOG10(@timedifference)/LOG10(@noedits)*10
END

-- Calculate the result value
SET @Result =
        (
            (
                (
                    (((@homerate * 0.25) +
                    (@schoolrate * 0.25) + (@expertise * 0.35) + (@age * 0.15)) * 0.30) +
                    (((@scale * 0.60) +
                    (@corners * 0.20) + (@anglemeasure * 20)) * 0.60) +
                    (@linusLaw * 0.10)
                )/10
            );

-- Return the result of the function
RETURN @Result;

END

```



```

GO
/***** Object: UserDefinedFunction [dbo].[fnVGTrust_v2]  Script Date:
2/07/2014 10:09:10 a.m. *****/
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO

CREATE FUNCTION [dbo].[fnVGTrust_v2]
(
    @featureX float,
    @featureY float,
    @homeX float,
    @homeY float,
    @schoolX float,
    @schoolY float,
    @subject varchar(50),
    @year varchar(7),
    @zoomscale int,
    --@vertexcount int,
    @noedits int,
    @firstedit datetime,
    @lastedit datetime,
    @shape geometry
)
RETURNS float
AS
BEGIN
    DECLARE @Result float

    DECLARE @homerate float, @schoolrate float, @expertise float, @age
float, @scale float, @corners float, @linusLaw float, @ratio float, @anglemeasure
float, @timedifference bigint;

    -- Calculate the home rating factor
    SET @homerate = dbo.fnVGTrust_ActivitySpaceNE(@featureX, @featureY,
@homeX, @homeY);

    -- Calculate the school rating factor
    SET @schoolrate = dbo.fnVGTrust_ActivitySpaceNE(@featureX,
@featureY, @schoolX, @schoolY);

    -- Calculate the expertise factor
    SET @expertise = dbo.fnVGTrust_Expertise(@subject);

    -- Calculate the age factor
    SET @age = dbo.fnVGTrust_SchoolYear(@year);

```



```

-- Calculate the scale factor
SET @scale = dbo.fnVGTrust_ZoomLevel(@zoomscale);

-- Calculate the linus law factor
SET @linusLaw = dbo.fnVGTrust_LinusLaw(@noedits);

-- Calculate the edits ratio
SET @timedifference = DATEDIFF(minute,@firstedit,@lastedit);

-- Rates the number of vertices
--SET @corners = dbo.fnVGTrust_Vertices(@vertexcount);
SET @corners = dbo.fnVGTrust_Vertices(@shape.STNumPoints());

-- Rates the angles
SET @anglemeasure = [dbo].[fnVGTrust_VertexRating](@shape);

IF @noedits <= 1
BEGIN
    SET @ratio = 0; --dbo.fnVGTrust_Vertices(@vertexcount);
END
ELSE
BEGIN
    SET @ratio = LOG10(@timedifference)/LOG10(@noedits)*10
END

-- Calculate the result value
SET @Result = (
    (
        (((@homerate * 0.25) + (@schoolrate * 0.25) + (@expertise
* 0.35) + (@age * 0.15)) * 0.40) +
        (((@scale * 0.60) + (@corners * 0.20) + (@anglemeasure *
0.20)) * 0.55) +
        (((@ratio * 0.4) + (@linusLaw * 0.6)) * 0.05)
    ) / 10
);

-- Return the result of the function
RETURN @Result;

END

GO

/***** Object: UserDefinedFunction [dbo].[fnVGTrust_ActivitySpaceLL]  Script
Date: 2/07/2014 10:09:10 a.m. *****/
SET ANSI_NULLS ON
GO

```



```

SET QUOTED_IDENTIFIER ON
GO
-- =====
-- Description: Returns a trust rating between 1 and 10 based on the features
euclidean distance from the contributors activity space. Can be used for more
than one activity space. Based in kilometres on latitude and longitude
coordinates. Classification is suitable for city scale, can be altered depending on
each case study.
-- =====
CREATE FUNCTION [dbo].[fnVGTrust_ActivitySpaceLL]
(
    @featureX float,
    @featureY float,
    @creatorX float,
    @creatorY float
)
RETURNS int
AS
BEGIN
    -- Declare the return variable here
    DECLARE @Result int, @r float = 6373.0, @lat1 float, @lon1 float, @lat2
float, @lon2 float, @dlon float, @dlat float, @a float, @c float, @distance float;

    SET @lat1 = RADIANS(@featureX);
    SET @lon1 = RADIANS(@featureY);
    SET @lat2 = RADIANS(@creatorX);
    SET @lon2 = RADIANS(@creatorY);

    SET @dlon = @lon2 - @lon1;
    SET @dlat = @lat2 - @lat1;
    SET @a = SQUARE(SIN(@dlat/2)) +
COS(@lat1)*COS(@lat2)*SQUARE(SIN(@dlon/2));
    SET @c = ATN2(SQRT(@a),SQRT(1-@a));
    SET @distance = @r * @c;

    SELECT @Result = CASE
        WHEN @distance <=1 THEN 10
        WHEN @distance > 1 AND @distance <= 2 THEN 9
        WHEN @distance > 2 AND @distance <= 5 THEN 8
        WHEN @distance > 5 AND @distance <= 10 THEN 7
        WHEN @distance > 10 AND @distance <= 20 THEN 6
        WHEN @distance > 20 AND @distance <= 30 THEN 5
        WHEN @distance > 30 AND @distance <= 50 THEN 4
        WHEN @distance > 50 AND @distance <= 100 THEN 3
        WHEN @distance > 100 AND @distance <= 200 THEN 2
        ELSE 1
    END
END

```

```

-- Return the result of the function
RETURN @Result

END

GO
/***** Object: UserDefinedFunction [dbo].[fnVGTrust_ActivitySpaceNE]  Script
Date: 2/07/2014 10:09:10 a.m. *****/
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO

-- =====
-- Description: Returns a trust rating between 1 and 10 based on the features
euclidean distance from the contributors activity space. Can be used for more
than one activity space. Based in metres on Northing and Easting coordinates.
Classification is suitable for city wide scale, can be altered depending on each case
study.
-- =====
CREATE FUNCTION [dbo].[fnVGTrust_ActivitySpaceNE]
(
    @featureX float,
    @featureY float,
    @creatorX float,
    @creatorY float
)
RETURNS int
AS
BEGIN
    -- Declare the return variable here
    DECLARE @Result int, @distance float;

    SET @distance = SQRT( SQUARE(@creatorX - @featureX) +
        SQUARE(@creatorY - @featureY));

    SELECT @Result = CASE
        WHEN @distance <=1000 THEN 10
        WHEN @distance > 1000 AND @distance <= 5000 THEN 9
        WHEN @distance > 5000 AND @distance <= 7500 THEN 8
        WHEN @distance > 7500 AND @distance <= 12500 THEN 7
        WHEN @distance > 12500 AND @distance <= 15000 THEN 6
        WHEN @distance > 15000 AND @distance <= 20000 THEN 5
        WHEN @distance > 20000 AND @distance <= 25000 THEN 4
        WHEN @distance > 25000 AND @distance <= 30000 THEN 3
        WHEN @distance > 30000 AND @distance <= 40000 THEN 2
        ELSE 1

```

```

END

-- Return the result of the function
RETURN @Result

END

GO
/***** Object: UserDefinedFunction [dbo].[fnVGTrust_CalcalateAngle]  Script
Date: 2/07/2014 10:09:10 a.m. *****/
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO
-- =====
-- Description: Calculates the angle of a point feature
-- =====
CREATE FUNCTION [dbo].[fnVGTrust_CalcalateAngle]
(
    @ptx float,
    @pty float
)
RETURNS float
AS
BEGIN
    -- Declare the return variable here
    DECLARE @Result float = null;

    IF @ptx = 0
    BEGIN
        IF @pty > 0
        BEGIN
            SET @Result = 90.0;
        END
        ELSE
        BEGIN
            SET @Result = 270.0;
        END
    END

    IF @pty = 0
    BEGIN
        IF @ptx > 0
        BEGIN
            SET @Result = 0.0;
        END
    END

```



```

        ELSE
        BEGIN
            SET @Result = 180.0;
        END
    END

    IF @Result IS NULL
    BEGIN
        SET @Result = ATAN(@pty/@ptx) * 180 / PI();

        IF @ptx < 0
        BEGIN
            SET @Result = @Result + 180;
        END
        ELSE
        BEGIN
            IF @pty < 0
            BEGIN
                SET @Result = @Result + 360;
            END
        END
    END

    -- Return the result of the function
    RETURN @Result

END

GO
/***** Object: UserDefinedFunction [dbo].[fnVGTrust_CalculateBearingRating]
Script Date: 2/07/2014 10:09:10 a.m. *****/
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO

CREATE FUNCTION [dbo].[fnVGTrust_CalculateBearingRating]
(
    @bearing1 float,
    @bearing2 float
)
RETURNS int
AS
BEGIN
    -- Declare the return variable here
    DECLARE @Result int, @diff float = @bearing2 - @bearing1;
    IF @diff < 0
    BEGIN

```




```

        SET @diff = @diff + 360;
    END

    SELECT @Result = CASE
        WHEN @diff > 85 AND @diff <= 95 THEN 10
        WHEN @diff > 175 AND @diff <= 185 THEN 9
        WHEN @diff > 115 AND @diff <= 125 THEN 8
        WHEN @diff > 125 AND @diff <= 135 THEN 6
        WHEN @diff > 265 AND @diff <= 275 THEN 7
        WHEN @diff > 275 AND @diff <= 359 THEN 5
        WHEN @diff > 185 AND @diff <= 195 THEN 4
        WHEN @diff > 45 AND @diff <= 60 THEN 3
        WHEN @diff <= 45 THEN 1
        ELSE 2
    END

    -- Return the result of the function
    RETURN @Result

END

GO
/***** Object: UserDefinedFunction [dbo].[fnVGTrust_CalculateDelta]  Script
Date: 2/07/2014 10:09:10 a.m. *****/
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO

CREATE FUNCTION [dbo].[fnVGTrust_CalculateDelta]
(
    @pt geometry,
    @compareWithPt geometry
)
RETURNS geometry
AS
BEGIN
    -- Declare the return variable here
    DECLARE @Result geometry

    -- Check the geometry types
    IF (@pt.STGeometryType() = 'Point' AND
    @compareWithPt.STGeometryType() = 'Point')
    BEGIN
        SET @Result = geometry::Point(@compareWithPt.STX -
    @pt.STX,@compareWithPt.STY - @pt.STY,2193);
    END

```



```

-- Return the result of the function
RETURN @Result

END

GO
/***** Object: UserDefinedFunction [dbo].[fnVGTrust_Expertise]  Script Date:
2/07/2014 10:09:10 a.m. *****/
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO

-- =====
-- Description: Classifies expertise based on school subject, and returns a higher
trust rating if subject is one associated with greater spatial ability. Expertise in
decending order from aList to dList. Returns a trust rating between 1 and 10.
-- =====
CREATE FUNCTION [dbo].[fnVGTrust_Expertise]
(
    @subject varchar(50)
)
RETURNS float
AS
BEGIN
    DECLARE @Result float

    -- Set the subject text to upper case for comparison purposes.
    SET @subject = UPPER(@subject);

    SELECT @Result = CASE
        WHEN @subject IN ('MATHEMATICS', 'SCIENCE', 'GRAPHICS',
'PHYSICS', 'CALCULUS', 'STATISTICS') THEN 10
        WHEN @subject IN ('SOCIALSTUDIES', 'ART', 'CHEMISTRY',
'BIOLOGY', 'GEOGRAPHY') THEN 7.5
        WHEN @subject IN ('ENGLISH', 'LANGUAGES', 'MATERIALS
TECHNOLOGY', 'HISTORY', 'ECONOMICS', 'ACCOUNTING', 'BUSINESS STUDIES',
'CLASSICS') THEN 5
        ELSE 2.5
    END

    -- Return the result of the function
    RETURN @Result

END

GO

```

```

/***** Object: UserDefinedFunction [dbo].[fnVGTrust_LinusLaw]  Script Date:
2/07/2014 10:09:10 a.m. *****/
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO

-- =====
-- Description: Returns a trust rating between 1 and 10 based on Linus' Law. This
reflects that the greatest improvement to a VG feature is made in the first six
contributions, with minor changes being made up to thirteen contributions. No
noticeable change is made by contributions in excess of thirteen edits.
-- =====
CREATE FUNCTION [dbo].[fnVGTrust_LinusLaw]
(
    @noedits int
)
RETURNS float
AS
BEGIN
    -- Declare the return variable here
    DECLARE @Result float

    SELECT @Result = CASE
        WHEN @noedits >= 13 THEN 10
        WHEN @noedits = 12 THEN 9
        WHEN @noedits = 11 THEN 9.5
        WHEN @noedits = 10 THEN 8
        WHEN @noedits = 9 THEN 7.6
        WHEN @noedits = 8 THEN 7.4
        WHEN @noedits = 7 THEN 7.2
        WHEN @noedits = 6 THEN 7
        WHEN @noedits = 5 THEN 6
        WHEN @noedits = 4 THEN 5
        WHEN @noedits = 3 THEN 4
        WHEN @noedits = 2 THEN 3
        WHEN @noedits = 1 THEN 2
        ELSE 1
    END

    -- Return the result of the function
    RETURN @Result
END

GO

```



```

/***** Object: UserDefinedFunction [dbo].[fnVGTrust_SchoolYear]  Script
Date: 2/07/2014 10:09:10 a.m. *****/
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO

```

-- Description: Returns a rating between 1 and 10 based on experience - ie school year. Written for application in a school context, this could also be classified by age bracket for wider public use. The trust rating is a linear function of school year. Forms an aspect of expertise rating of the contributor.

-- =====

```

CREATE FUNCTION [dbo].[fnVGTrust_SchoolYear]
(
    @year varchar(7)
)
RETURNS int
AS
BEGIN
    DECLARE @Result int;

    -- Set the year text to upper case for comparison purposes.
    SET @year = UPPER(@year);

    SELECT
        @Result = CASE
            WHEN @year = 'YEAR 13' THEN 10
            WHEN @year = 'STAFF' THEN 10
            WHEN @year = 'YEAR 12' THEN 8
            WHEN @year = 'YEAR 11' THEN 6
            WHEN @year = 'YEAR 10' THEN 4
            WHEN @year = 'YEAR 9' THEN 3
            ELSE 2
        END

    -- Return the result of the function
    RETURN @Result

```

END

```

GO
/***** Object: UserDefinedFunction [dbo].[fnVGTrust_UserRating]  Script
Date: 2/07/2014 10:09:10 a.m. *****/
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON

```

GO

```
-- =====
-- Description: returns a total weighting for user trust based on activity space and
-- expertise. The below function has been written accounting for two activity spaces
-- - home and work, or in this case, school. Expertise is reflected also for school
-- students, based on their year at school (experience), and their best subject. This
-- could be altered based on the individual case study. In total activity space
-- accounts for 50% of the user trust rating, with expertise accounting for the other
-- 50%.
```

```
-- =====
CREATE FUNCTION [dbo].[fnVGTrust_UserRating]
(
    @schooldist float,
    @homedist float,
    @year float,
    @subject float
)
RETURNS float
AS
BEGIN
    DECLARE @Result float

    -- Add the T-SQL statements to compute the return value here
    SELECT @Result = ((@schooldist * 0.25) + (@homedist * 0.25) + (@year *
0.15) + (@subject * 0.35))

    -- Return the result of the function
    RETURN @Result

END
```

GO

```
/***** Object: UserDefinedFunction [dbo].[fnVGTrust_VertexRating] Script
Date: 2/07/2014 10:09:10 a.m. *****/
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO
```

```
-- =====
-- Description: Calculates the rating value for a shape based on its geometry and
-- the number of vertices and angles bewteen them
```

```
-- =====
CREATE FUNCTION [dbo].[fnVGTrust_VertexRating]
(
```

```

        @polygon geometry
    )
    RETURNS float
    AS
    BEGIN
        -- Declare the return variable here
        DECLARE @Result float;

        -- Get the number of points in the supplied shape
        DECLARE @i int = 1, @cnt int, @pointA geometry, @pointB geometry,
        @delta geometry, @bearing float, @bearing2 float, @rating float = 0;
        DECLARE @bearings TABLE (ID int PRIMARY KEY IDENTITY(1,1),
        bearing float);
        DECLARE @ratings TABLE (ID int PRIMARY KEY IDENTITY(1,1), rating
        float);

        DECLARE @bea varchar(MAX) = '';
        DECLARE @rat varchar(MAX) = '';

        SET @cnt = @polygon.STNumPoints();

        -- Iterate through each point in the shape to generate the bearings list
        WHILE @i <= @cnt
        BEGIN
            -- Get the point for comparison
            SET @pointA = @polygon.STPointN(@i);

            IF @i = @cnt
            BEGIN
                SET @pointB = @polygon.STPointN(1);
            END
            ELSE
            BEGIN
                SET @pointB = @polygon.STPointN(@i + 1);
            END

            -- Calculate the delta value
            SET @delta = dbo.fnVGTrust_CalculateDelta(@pointA, @pointB);

            IF @delta.STX <> 0 AND @delta.STY <> 0
            BEGIN
                -- Calculate the bearing
                SET @bearing =
                [dbo].[fnVGTrust_CalculateAngle](@delta.STX, @delta.STY);
                INSERT INTO @bearings(bearing) values(@bearing);

                SET @bea = @bea + CAST(@bearing as varchar(20)) + ',';
            END
        END
    
```



```

-- Iterate to the next point
SET @i = @i + 1;
END

-- Insert the first bearing again
INSERT INTO @bearings(bearing)
SELECT bearing from @bearings WHERE ID = 1;

-- Iterate through each bearing in the list
SELECT @cnt = MAX(ID) FROM @bearings;

SET @i = 1;
WHILE @i < @cnt
BEGIN
    SELECT @bearing = bearing FROM @bearings WHERE ID = @i;
    SELECT @bearing2 = bearing FROM @bearings WHERE ID = @i + 1;

    SET @rating = [dbo].[fnVGTrust_CalculateBearingRating](@bearing,
@bearing2);

    INSERT INTO @ratings(rating) VALUES (@rating);

    SET @rat = @rat + CAST(@rating as varchar(20)) + ',';

    -- Iterate to the next point
    SET @i = @i + 1;
END

-- Set the rating as the sum of the individual ratings
SELECT @Result = SUM(rating)/COUNT(rating) FROM @ratings;

-- Return the result of the function
RETURN @Result

END

GO
/***** Object: UserDefinedFunction [dbo].[fnVGTrust_Vertices]  Script Date:
2/07/2014 10:09:10 a.m. *****/
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO

```

```

-- =====
-- Description: Returns a rating between 1 and 10 for the zoom level at which a
feature is captured. The function below is written to reflect esri zoom parameters,

```



ie: 1:250, 1:500 etc. This would need to be altered to suit other applications such as Google Maps, where zoom is classified as an integer between 1 and 17.

```
-- =====
CREATE FUNCTION [dbo].[fnVGTrust_Vertices]
(
    @VertexCount int
)
RETURNS int
AS
BEGIN
    DECLARE @Result int

    SELECT @Result = CASE
        WHEN @VertexCount = 3 THEN 1
        WHEN @VertexCount = 4 THEN 5
        WHEN @VertexCount > 4 AND @VertexCount <= 6 THEN 6
        WHEN @VertexCount > 6 AND @VertexCount <= 8 THEN 7
        WHEN @VertexCount > 8 AND @VertexCount <= 10 THEN 8
        WHEN @VertexCount > 10 AND @VertexCount <= 12 THEN 9
        ELSE 10

    END

    -- Return the result of the function
    RETURN @Result

END

GO

/***** Object: UserDefinedFunction [dbo].[fnVGTrust_ZoomLevel]  Script
Date: 2/07/2014 10:09:10 a.m. *****/
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO

-- =====
-- Description: Returns a rating between 1 and 10 for the zoom level at which a
feature is captured. The function below is written to reflect esri zoom parameters,
ie: 1:250, 1:500 etc. This would need to be altered to suit other applications such
as Google Maps, where zoom is classified as an integer between 1 and 17.
-- =====
CREATE FUNCTION [dbo].[fnVGTrust_ZoomLevel]
(
    @zoomscale int
```

```

)
RETURNS int
AS
BEGIN
    DECLARE @Result int

    SELECT @Result = CASE

        IF @VertexCount <= 3
        BEGIN
            SET @Result = 0
        END

        ELSE IF @anglemeasure <= 3
        BEGIN
            SET @Result = @anglemeasure
        END

        ELSE SELECT @Result = CASE

            WHEN @zoomscale = 250 THEN 10
            WHEN @zoomscale = 500 THEN 8
            WHEN @zoomscale = 1000 THEN 6
            WHEN @zoomscale = 2000 THEN 5
            WHEN @zoomscale = 4000 THEN 4
            WHEN @zoomscale = 8000 THEN 3
            WHEN @zoomscale = 16000 THEN 2
            WHEN @zoomscale = 32000 THEN 1
            ELSE 0

        END

    END

    -- Return the result of the function
    RETURN @Result

END

GO

/* -----
--
TRIGGER FUNCTIONS ON LAYER TABLE FIRED WHEN A FEATURE IS INSERTED OR
UPDATED
-----
*/

```

```

/***** Object: Trigger
[dbo].[trCROWDSOURCE_NZTM_BuildingFootprint_Insert]  Script Date:
2/07/2014 10:11:39 a.m. *****/
SET ANSI_NULLS ON
GO

SET QUOTED_IDENTIFIER ON
GO

-- =====
-- Description: Updates the geometry statistics and calculates the trust details for
the captured buildings when a new feature is created
-- =====
CREATE TRIGGER [dbo].[trCROWDSOURCE_NZTM_BuildingFootprint_Insert]
ON [dbo].[CROWDSOURCE_NZTM_BUILDINGFOOTPRINT]
AFTER INSERT
AS
BEGIN
    -- SET NOCOUNT ON added to prevent extra result sets from
    -- interfering with SELECT statements.
    SET NOCOUNT ON;

    UPDATE dbo.CROWDSOURCE_NZTM_BuildingFootprint
    SET
        NZTMX = i.SHAPE.STCentroid().STX,
        NZTMY = i.SHAPE.STCentroid().STY,
        NUMBER_OF_EDITS = 1,
        GEOM_AREA = i.SHAPE.STArea(),
        ACCURACY_CLASS = dbo.fnVGTrust(i.SHAPE.STCentroid().STX,
i.SHAPE.STCentroid().STY, sb.NZTMX, sb.NZTMY, sc.NZTMX, sc.NZTMY,
u.[FavouriteSubject], u.[ClassYear], i.CAPTURE_SCALE, 1, i.[CAPTURED_DATE],
i.[DATE_LAST_CHANGE], i.Shape)
    FROM
        dbo.CROWDSOURCE_NZTM_BuildingFootprint F

    INNER JOIN
        inserted i ON
            F.OBJECTID = i.OBJECTID

    LEFT JOIN
        [GISPublicViewerSettings].[dbo].[AspNetUsers] U ON
            i.[CAPTURED_BY] = U.UserName

    LEFT JOIN
        [dbo].[CROWDSOURCE_NZTM_SCHOOLS] sc ON
            U.[School] = sc.[NAME]

```



```

LEFT JOIN
[dbo].[CROWDSOURCE_NZTM_SUBURBS] sb ON
    U.[Suburb] = sb.[SUBURB];

-- Insert a record in the change log
INSERT INTO
dbo.CROWDSOURCE_NZTM_BUILDINGFOOTPRINT_ChangeLog(GlobalID,
BLD_CLASS, BLD_OCCUPATON, BLD_FLOORS, CAPTURE_SCALE, GEOM_AREA,
ACCURACY_CLASS, CHANGE_BY, CHANGE_DATE, SHAPE, COMMENTS)
SELECT
    F.GlobalID, F.BLD_CLASS, F.BLD_OCCUPATON, F.BLD_FLOORS,
    F.CAPTURE_SCALE, F.GEOM_AREA, F.ACCURACY_CLASS, F.[CAPTURED_BY],
    GETDATE(), F.SHAPE, F.Comments
FROM
    inserted i

INNER JOIN
    dbo.CROWDSOURCE_NZTM_BuildingFootprint F ON
        F.OBJECTID = i.OBJECTID;

END

GO

/***** Object: Trigger
[dbo].[trCROWDSOURCE_NZTM_BuildingFootprint_Update]  Script Date:
2/07/2014 10:12:59 a.m. *****/
SET ANSI_NULLS ON
GO

SET QUOTED_IDENTIFIER ON
GO

-- =====
-- Description: Updates the geometry statistics and calculates the trust details for
the captured buildings when an existing feature is updated
-- =====
CREATE TRIGGER [dbo].[trCROWDSOURCE_NZTM_BuildingFootprint_Update]
    ON [dbo].[CROWDSOURCE_NZTM_BUILDINGFOOTPRINT]
    AFTER UPDATE
AS
BEGIN
    -- SET NOCOUNT ON added to prevent extra result sets from
    -- interfering with SELECT statements.
    SET NOCOUNT ON;

    IF UPDATE(SHAPE)

```



```

BEGIN
    UPDATE dbo.CROWDSOURCE_NZTM_BuildingFootprint
    SET
        NZTMX = i.SHAPE.STCentroid().STX,
        NZTMY = i.SHAPE.STCentroid().STY,
        NUMBER_OF_EDITS = ISNULL(d.NUMBER_OF_EDITS,1) + 1,
        GEOM_AREA = i.SHAPE.STArea()
        ,ACCURACY_CLASS =
        dbo.fnVGTrust(i.SHAPE.STCentroid().STX, i.SHAPE.STCentroid().STY, sb.NZTMX,
        sb.NZTMY, sc.NZTMX, sc.NZTMY, u.[FavouriteSubject], u.[ClassYear],
        i.CAPTURE_SCALE, ISNULL(d.NUMBER_OF_EDITS,1) + 1, i.[CAPTURED_DATE],
        i.[DATE_LAST_CHANGE], i.Shape)

    FROM
        dbo.CROWDSOURCE_NZTM_BuildingFootprint F

        INNER JOIN
        inserted i ON
            F.OBJECTID = i.OBJECTID

        INNER JOIN
        deleted d ON
            D.OBJECTID = d.OBJECTID

        LEFT JOIN
        [GISPublicViewerSettings].[dbo].[AspNetUsers] U ON
            i.[CAPTURED_BY] = U.UserName

        LEFT JOIN
        [dbo].[CROWDSOURCE_NZTM_SCHOOLS] sc ON
            U.[School] = sc.[NAME]

        LEFT JOIN
        [dbo].[CROWDSOURCE_NZTM_SUBURBS] sb ON
            U.[Suburb] = sb.[SUBURB];

    -- Insert a record in the change log
    INSERT INTO
        dbo.CROWDSOURCE_NZTM_BUILDINGFOOTPRINT_ChangeLog(GlobalID,
        BLD_CLASS, BLD_OCCUPATON, BLD_FLOORS, CAPTURE_SCALE, GEOM_AREA,
        ACCURACY_CLASS, CHANGE_BY, CHANGE_DATE, SHAPE, COMMENTS)
    SELECT
        F.GlobalID, F.BLD_CLASS, F.BLD_OCCUPATON,
        F.BLD_FLOORS, F.CAPTURE_SCALE, F.GEOM_AREA, F.ACCURACY_CLASS,
        F.[CAPTURED_BY], GETDATE(), F.SHAPE, F.Comments
    FROM
        inserted i

```



```
INNER JOIN
dbo.CROWDSOURCE_NZTM_BuildingFootprint F ON
    F.OBJECTID = i.OBJECTID;
END
END
GO
```